Department for Business, Energy & Industrial Strategy

RESEARCH

BEIS RESEARCH PAPER NUMBER 4

Industrial Clusters in England

SEPTEMBER 2017

Contents

Executive Summary

NIESR, SpazioDati and City REDI were commissioned by the Department for Business, Innovation and Skills to generate new evidence on UK industrial clusters and to test the potentials and limitations of “big-data” techniques applied to the study of the topic.

The report showcases an innovative data-driven approach to investigate the patterns of geographical clustering and functional integration across three sectors: digital health, financial services and the processing industry. These three sectors represent an emerging industry, an established service sector and a manufacturing sector with the presence of a formal cluster organisation.

Quantitative analyses were complemented by qualitative case studies based on interviews with key stakeholders. Semi-structured questionnaires generated detailed information on the evolution of the three selected clusters and the nature of the relationships between companies with other companies and local institutions.

It is concluded that general features of this “big-data” methodology of industry classification can in general be applied to map clusters in emerging sectors not easily classified by the current Standard Industrial Classification (SIC) system. However, a number of industry specific issues require ad-hoc solutions.

Quantitative Analysis

Internet data were used to identify companies belonging to each of the sectors being studied. Proprietary tools were used to screen and collect relevant information from companies’ websites, including geographical location, concepts describing a company’s activities and its web-links to other institutions.

An algorithm was applied to identify clusters based on the physical distance between companies. This approach allowed companies to be classified as part of the same sector even when they were classed with different SIC codes, and could identify clusters spreading over multiple discrete administrative areas.

This approach revealed similarities and differences in the patterns of geographical agglomeration across the three sectors. The largest urban areas emerged as important agglomeration areas for all three sectors. For example, London, Birmingham and Manchester were consistently identified as the largest sectoral agglomerations. Smaller urban areas had a different importance across sectors.

It is suggested that companies in these sectors are attracted to large metropolitan areas by factors that are common to the larger population of UK companies. These factors included the proximity with larger product and labour markets, and access to strategic tangible and intangible infrastructures within larger cities.

More stringent criteria to identify clusters were used to control for these factors. However, many clusters continued to be identified with the more stringent approach. This was

interpreted as evidence of positive externalities from the co-location of similar companies within a specific geographical area.

The work demonstrated how important differences in functional relationships between companies and institutions between sectors may exist. Analysis of the network of web-links extracted from companies' webpages found that digital health and processing industry companies' website contained frequent links to the websites of academic institutions. By contrast, financial services companies frequently linked to the same government websites common with other companies.

The study also identified the possible influence of sector-specific factors for digital-health. Oxford and Cambridge emerge as the only geographical areas where their relative concentration is at least two times the national average. No similar locations were identified for the remaining two sectors.

Qualitative Analysis

Qualitative data was used to gain insights into the strategic importance of the relationships developed by a company with other organizations inside and outside clusters. In addition, interviewers explored how companies' officials perceived the value of locating within an industry cluster and their experience of the opportunities generated by geographical agglomeration and networking.

Some clusters are located where they are for historical reasons. The case studies on the financial services sector in Leeds and on the North East of England Process Industry Cluster (NEPIC) organisation in Teesside Valley emphasized the important role of historical legacy and central organisation in the establishment of these clusters. These two case studies shed light on the benefits arising from co-location of companies in the same industry or closely integrated industries.

The case study exploring digital health companies in Birmingham revealed that only six of the ten companies classified at digital health based on website data related to a strict definition of the sector. The other cases were generally pharmaceutical companies. In line with the limited number of inter-company web-links for this sector, the case study suggested that the agglomeration of digital health companies in Birmingham is not generally perceived as a functional cluster, and that there are no significant partnerships between the companies in the area.

All case studies confirmed the differing role of company-university relationships between sectors inferred by comparing network graphs of web-links. While companies in the NEPIC processing industry cluster and the digital health sector report strategic relationships with universities, this is not the case for financial companies.

The following report has undergone some minor editing prior to publication.

Industrial Clusters in England

National Institute of Economic and Social Research

SpazioDati

City REDI (University of Birmingham)

Industrial Clusters in England

By:

NIESR	SpazioDati	City REDI
Michele Bernini	Michele Barbera	Simon Adderley
Rebecca Riley	Andrey Bratus	Rachel Mulhall
Ana Rincón Aznar	Nicola Sambin	Paulina Ramirez

With a critical appraisal by

Max Nathan (City REDI)

1. Introduction

Industrial clusters are regarded as an interesting laboratory for economic research and a useful concept to guide industrial and labour-market policies (Porter, 2000). The indefiniteness of cluster boundaries generates interesting opportunities for empirical exploration, both from a geographical and an industry perspective, but the formulation of policy objectives may require constraining cluster definitions to match the boundaries of administrative units, of particular sets of industries, or formal associations of enterprises and institutions (e.g., COSME EU Programme). This report explores the possibility of using Internet data to identify industrial clusters. Qualitative case studies complement and test the innovative quantitative approach.

Michael Porter provides a definition of clusters that hinges on three key aspects: the geographical location of companies (i.e., co-location), their functional relationships (i.e., supply chain relationships, production of complementary products), and the presence of institutional linkages (i.e., institutional of formal associations or special linkages with local authorities and universities) (Porter, 1998). The trade-off between operational and analytical objectives is reflected in each of the three dimensions of Porter's definition.

On the geographical dimension, previous work undertaken by the Enterprise Research Centre (ERC) for the UK Department for Business, Innovation and Skills (BIS) exploits the boundaries of individual Local Enterprise Partnerships (LEP) to identify the concentration (in terms of employment) of 11 strategic industries or 5-digit Standard Industrial Classification (SIC) activities (Anyadike-Danes et al. 2013). While that type of analysis is useful for identifying specific local comparative advantages across England's Local Enterprise Partnerships (LEPs) (BIS Research Strategy 2014-2015), it overlooks important information on clusters spanning across different LEPs or involving companies classified under different SIC categories. The work of Duranton and Overman (2005, 2008) departs from the use of discrete geographical units and locates UK companies in a continuous geographical space using establishments' post-codes reported in the Annual Respondent Database (ARD) and the Code-Point dataset. In these works, localisation indicators at the industry-level are constructed by estimating probability densities of distances between pairs of establishments belonging to the same industry.

Because industry-level analyses do not require one to establish ex-ante a particular geographical segmentation, they are more appropriate to identify patterns of firm co-location that are not confined within the boundaries of discrete geographical units (Simpson, 2007). However, the focus of these studies on spatial relationships within the same industry class (i.e., classified under the same SIC code) overlooks key inter-industry functional relationships along the supply chain. As a consequence, industry-level studies capture only some of the functional interdependencies between companies that generate agglomeration economies. A US-based study by Delgado, Porter and Stern (2014) proposes a data-driven methodology based on the co-location of establishments from different industries in the same region to identify functional relationships across industries. Their proposed algorithm identifies 'clusters' of industries that are more likely to give rise to co-location of companies.

A merit of the data-driven approach is that it does not rely on any specific geographical unit of analysis, and it can be used to compare clusters across different countries. The

main limitation of existing work is that it assumes that standard industry classifications are appropriate to represent the nature of economic activities. Classifications may be outdated and may not accurately describe the essence of innovative activities. This problem is especially significant when trying to classify the Digital Economy. A NIESR report commissioned by Google exploits information collected from companies' websites to identify UK 'digital' businesses (Nathan and Rosso, 2013). This research demonstrates that the use of standard SIC codes leads to underestimation of the size of the UK digital economy, as many of the businesses producing digital products are not captured by the relevant SIC codes.

Porter's definition of clusters encompasses the institutional ties between geographically proximate and functionally integrated companies or between companies and other stakeholders. Some of the UK's industrial clusters are supported by agencies; for example, Tech City UK was established in 2010 to lead the development of the East London technology cluster. The North East Process Industry Cluster (NEPIC) is a second example of an organisation formed by companies that operate in interrelated sectors (i.e., chemicals, polymers, pharmaceuticals, biotechnology, and renewables) and that are co-located in the North East of England. On the one hand, these formal associations respond to policy inputs such as the UK ministers' renewed interest in adopting an 'industrial strategy' (Nathan and Vandore, 2014); on the other hand they signal the existence of agglomeration economies, the demand for specialized governance bodies, and the need for shared tangible and intangible infrastructures.

From an empirical perspective, these formal associations are useful sources of information to identify the key private and public stakeholders in a cluster. However, supply chain relationships and institutional linkages may extend well beyond these organisations. For instance, individual companies may form partnerships with universities and other public institutions that are not formal members of a cluster organisation. In addition, not all clusters may have reached the same level of maturity, or have expressed the need to establish more formal partnerships.

This report uses a novel data-driven approach, complemented by qualitative analysis, to investigate the patterns of geographical clustering and functional integration across three sectors: Digital-Health, Financial Services and Processing Industry. These three sectors have been selected by BIS to represent respectively an emerging industry, an established service sector and a manufacturing sector with the presence of a formal cluster organisation. The first of the three sectors is an emerging industry including manufacturing and services companies that develop or apply new technologies to the field of Human Health. Since this is an emerging sector encompassing different economic activities, it cannot easily be mapped into SIC codes. The Financial Services industry instead is an established sector that can be more easily mapped into SIC codes. The Processing Industry includes companies from different SIC classes such as Chemicals, Pharmaceuticals, and Engineering and at the local level is strongly integrated with service companies.

We combine quantitative analyses based on unstructured Internet data and qualitative case studies based on interviews with key stakeholders. The quantitative analyses are conducted by NIESR and SpazioDati. SpazioDati is based in Trento (Italy) and operates at the crossroad between Big Data and Semantic Web. SpazioDati's proprietary tools are used to screen and collect relevant information from companies' websites, such as

concepts describing a company's activity and its web-links to other institutions. Based on this information, NIESR and SpazioDati have worked closely together to classify companies into the three sectors under investigation. Density based clustering algorithms have then been used to identify the patterns of geographical agglomeration of these companies. In this respect, the quantitative analysis departs considerably from previous work on UK clusters. First, by using Internet data to identify which companies belong to each sector our approach is mostly data-driven and it is less reliant on SIC codes. This allows us to classify as part of a sector companies with different SIC codes. Second, by using an algorithm that identifies clusters based on the physical distance between companies we can identify clusters spreading over multiple discrete administrative areas.[1] Third, by investigating the links included on companies' websites we can investigates the relationships between companies and between companies and other institutions.

The qualitative analysis is conducted by City REDI, a recently established research institute within the Birmingham Business School working to develop the academic understanding of city regions across the globe. Based on a common semi-structured questionnaire, qualitative case studies are conducted to generate detailed information on the evolution of specific clusters (i.e., one for each sector under analysis), on the nature of the relationships between companies and between companies and local institutions. Qualitative data is also used to gain insights into the strategic importance of the relationships developed by a company with other organizations inside and outside the cluster. During interviews, we also explore how companies' officials perceive the value of locating within an industry cluster and what is their experience of the opportunities generated by geographical agglomeration and networking.

We include at the end of this report a critical appraisal of the project written by Max Nathan, a Senior Fellow at the Birmingham Business School and a Deputy Director of the What Works Centre for Local Economic Growth. This appraisal was written after the completion of the research and it provides valuable insights on the contribution of this study to the wider research programme on industrial clusters. By highlighting the limitations of the project, this appraisal also suggests areas for future research and how Big Data approaches can be more deeply integrated with qualitative methodologies.

[1] For simplicity of exposition we will name each cluster after the Travel-to-Work Area (TTWA) that includes most of its companies. However we allow for a cluster to extend across multiple TTWAs.

2. Quantitative analysis: methodology

This section describes the five main stages of the quantitative analysis:

1. Sample selection
2. Data collection
3. Company classification
4. Identification of geographical clusters
5. Link analysis

As a starting point we select a set of companies and collect unstructured textual data from their websites by using SpazioDati proprietary tools. These data are used to compute quantitative indicators that allow us to establish which companies belong to the sector under examination. We then use a density based clustering algorithm to identify geographical concentrations of companies and eventually investigate relationships between companies and institutions by using weblinks included within companies' websites.

2.1 Sample selection

The quantitative analysis hinges on two samples of firms. The ***extended sample*** is the sample of all companies for which we collect data from the Internet. This sample is defined irrespectively of the sector under analysis and it includes a large cross-section of UK companies registered with Company House. Company House data are obtained from the Financial Analysis Made Easy (FAME) database provided by Bureau Van Dijk. Four criteria guide the inclusion of companies within this sample:

a) The company must be active (i.e., exists as a legal entity) at December 2015 when the data are downloaded. We adopt a restrictive definition of active companies that excludes firms that are in receivership, that are dormant or in default. The exclusion of firms with a greater probability of death extends the longevity of our study, by making our results more robust to short term changes in the population of firms due to the exit of companies currently subject to administrative measures.

b) The website of the company must be reported in FAME. Bureau Van Dijk analysts collect website addresses for most of the companies in FAME. By conducting random checks on the data we conclude that missing addresses mostly arise when companies do not have a website.

c) The company must be either a limited liability company (i.e., Limited Liability Partnership, Limited Partnership, Private Limited, Unlimited) or a public company (i.e., Public AIM, Public Quoted OFEX, Public Not Quoted, Public Quoted). This criterion is needed to obtain a more homogeneous sample of for-profit companies.

d) The company must not have the same website as another company with a different Registration Office postcode. This restriction arises for two reasons. First, it is

problematic to associate specific online contents across companies that share a website. Second, when a website is shared by companies registered at different locations, it is not possible to exploit website contents to identify geographical clusters.

Table 1 reports the number of unique companies and unique websites retained in the extended sample after applying the selection criteria.[2] Selection rule (c) results in a slight reduction in the number of retained firms, while criterion (d) causes a large loss of about 180,000 companies and an 8% reduction in the number of unique websites. There is also evidence that criterion (d) affects the size composition of the sample, as the proportion of SMEs increases by about three percent after applying this filter.

Table 1: Composition of the extended sample

	Unique companies	Unique websites	% SME
Criteria (a) & (b)	742,354	624,202	92.4%
Criteria (a) & (b) & (c)	703,755	591,864	92.5%
Criteria (a) & (b) & (c) & (d)	545,661	545,661	95.8%

Notes: The table reports the number of unique companies, the number of unique websites and the percentage of SMEs. SME are defined according to the European Commission guidelines: they have less than 250 employees, turnover smaller than £42.3m or balance sheet smaller than £36.4m.

Differently from the extended sample, the composition of the ***restricted sample*** changes when we analyse different sectors. We include in this sample firms from the extended sample that on the basis of prior information we classify as belonging to the sector of interest. The information that we use to construct this sample varies across sectors.

For the Digital Health sector, the inclusion of firms in the restricted sample is based on a list of companies provided by the UK Office of Life Sciences (OLS). This is a list compiled by OLS officials on the basis of industry-specific knowledge. The restricted sample for the Processing Industry is based on a subset of companies formally associated with the North East Processing Industry Cluster organisation (NEPIC).[3] From the member list of this organisation we include only companies that strictly belong to the Processing Industry according to the SIC codes of their main economic activity.[4] This restriction is necessary because the list of NEPIC members includes companies conducting very different economic activities and their unrestricted inclusion in the restricted sample would generate a definition of the industry that is too general to be useful. This does not imply that the data-driven definition of the sector will exclude companies that are not strictly part of the Processing Industry. For example, a producer of machineries that are used by Chemical or Pharmaceutical companies is likely to be identified as part of the industry if its website describes the final use of its products or if it includes a description of their typical clients.

[2] For convenience, criteria (a) and (b) are imposed when downloading the data. This measure considerably reduces the time of downloading the data from the FAME website. For this reason we do not report the size of the sample before imposing these criteria.
[3] We obtain the list of NEPIC members from the website of the association (*www.nepic.co.uk*).
[4] These are firms whose main economic activity is classified with one of the following SIC codes: 20110, 20130, 20140, 20160, 20590, 21100, 21200, 46750. These codes broadly define the chemical/pharmaceutical sectors of the Processing industry.

For the financial sector, the restricted sample is constructed on the basis of SIC codes only. More specifically, we include in this sample only companies whose main economic activity is classified under SIC codes related to banking, insurance or auxiliary financial activities.[5] The use of different information sources to construct the sector-specific restricted samples provides an opportunity to evaluate which kind of prior knowledge is more appropriate to initialize industry-classification exercises based on website analysis. Table 2 describes the size and the information source for each one of the three sector-specific restricted samples. It is apparent that the selection criteria adopted to construct these samples lead to very different sample sizes, ranging from 53 companies for the Processing Industry to over 4,500 for the financial sector.

This striking difference in size reflects differences in the trade-off between precision in selecting the units to include in the sample and the representativeness of the sample. Because for the Digital-Health and the Processing Industry we use lists of companies respectively identified by the OLS or that are members of an industry association, we are more confident that the websites of these companies provide high quality information to characterize these sectors. On the contrary, when using SIC codes for the Financial Sector we capture a more inclusive sample of firms whose websites generate a noisier characterization of the sector.

Table 2: Composition of the restricted samples

Sector	Size (num. firms)	Information source
Digital-Health	378	OLS list
Processing Industry	53	NEPIC membership, SIC codes
Financial Sector	4,551	SIC codes

Notes: The table reports the numbers and the percentages of successfully indexed WebPages over the total number of unique webpages in the sample of companies.

2.2 Data collection

The websites of all companies in the extended sample are automatically uploaded and indexed on SpazioDati's servers by using a web crawler.[6] When a company's website is moved to an address different from the one reported in FAME, the crawler tracks the website and registers the new address. Table 3 reports the percentage of websites that are successfully indexed. In 85% of the cases the web crawler is successful in fetching and indexing a company's website.

[5] These are firms whose main economic activity is classified with one of the following SIC codes: 64191, 64192, 65110, 65120, 65201, 65202, 65300, 66110, 66120, 66190, 66210, 66220, 66290, 66300.
[6] A web crawler is a software application that runs automated tasks over a set of websites.

Table 3: Statistics on crawling

	Number of websites	% total
Successful indexing	530,016	85%
Failed indexing	94,186	15%
Websites total	624,202	

Notes: The table reports the numbers and the percentages of successfully indexed webpages over the total number of unique webpages in the sample of companies. Crawling is conducted on all companies' websites (i.e., filtering criteria a, b, c, d apply to the sample of firms retained for cluster analysis but not to the sample of websites fed into the algorithm to define an industry).

The proprietary textual analysis algorithm developed by SpazioDati recognizes and extracts ***entities*** from successfully indexed websites. Entities are combination of words that provide information on a company and on its lines of business. Important entities are distinguished from unimportant ones (i.e., noise) by jointly exploiting two attributes:

a) The centrality of an entity within the text. Centrality is measured by using a graph-based ranking algorithm inspired by TextRank.[7] The algorithm can be thought of as an application to textual analysis of the celebrated PageRank[8] algorithm used by Google to rank websites among search results. PageRank establishes the relevance of a website by taking into account the number of links pointing to it and the importance of the websites from which the links are departing. Similarly, we can infer the relevance of an entity within a text by jointly considering the frequency of that entity in the text and the frequency of the entities close to it (i.e., in the same sentence, in the same paragraph etc...). An entity that appears frequently in the text of a website, or that appear close to other frequent entities, is given a higher score.[9]

b) The position of the entity within the website's structure. Entities that appear in the most visible positions within a website (i.e., at the top of the homepage) are more likely to provide important information on a company's activity.

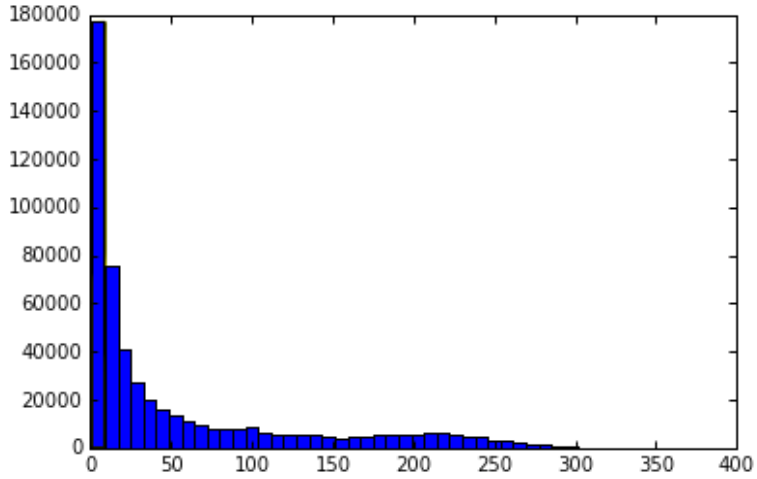
Websites with many pages may facilitate the identification of important entities as they provide a richer set of possible locations for individual entities. However, these websites are also more likely to describe a diversified business group rather than a unique company. In this case, it is difficult to extract precise information on individual subsidiaries' activities. Figure 1 shows the distribution of websites by number of pages. The shape of the histogram is typical of a power-law distribution with a strong accumulation point around low values (i.e., 88,717 websites have only one page) and a long tail to the right. We

[7] https://web.eecs.umich.edu/~mihalcea/papers/mihalcea.emnlp04.pdf
[8] https://en.wikipedia.org/wiki/PageRank

[9] More precisely, a text is transformed into a graph whose nodes are entities. In such a graph links are drawn among entities when these entities appear close to each other within the text. By doing so, we construct a representation of a text that is similar to the representation of the Internet as a network of connected webpages. The PageRank algorithm is then run on the graph representing a text to assign a "centrality" score to each entity and entities are sorted according to their scores. Eventually, we select from that text only the entities with the highest score and discard the remaining ones as noise.

interpret the accumulation of websites at the lower end of the distribution as favourable to reduce the risk of misclassification arising from business groups' websites.

Figure 1: Distribution of websites by number of pages



Notes: The figure represents the distribution of successfully indexed websites by the number of webpages.

SpazioDati's algorithm is iterated over individual websites. A selected group of entities from the websites of the companies in the restricted sample are used to generate a ***benchmark set of entities*** that characterizes the sector under investigation. Ideally, the benchmark set of entities represents companies that conduct similar economic activities, use similar technologies and that target similar markets. The quality of this set relies on the composition of the *restricted samples,* and in turn on the information exploited to generate these samples.[10]

We provide an illustration of how the benchmark set of entities is constructed by using the Digital-Health sector as an example. In Figure 2, we plot a selected group of entities extracted from the websites of companies in the Digital-Health restricted sample. The x-axis measures the frequency of the entity within the restricted sample (i.e., how many times the entity is extracted from the websites of companies in the restricted sample). This statistic is a naïve indicator of the importance of an entity within the sample because it gives too much weight to general entities that are likely to appear very frequently both in the restricted and in the extended sample. Hence, this statistic is insufficient to capture entities that truly characterize the sector of the companies in the restricted sample. For example, entities such as "product" or "services" are bound to appear very frequently across all sectors. For this reason they are not good candidates to represent a specific sector. To better identify entities that are specific of the Digital-Health sector we compute a normalized frequency score $nfr(t)$ (plotted on the y-axis) that rewards an entity's frequency in the restricted sample while penalizing its frequency in the extended sample:

$$nfr(t) = fr_{rs}(t) \cdot \ln\left(\frac{N}{fr_{es}(t)}\right)$$

[10] See Table 2.

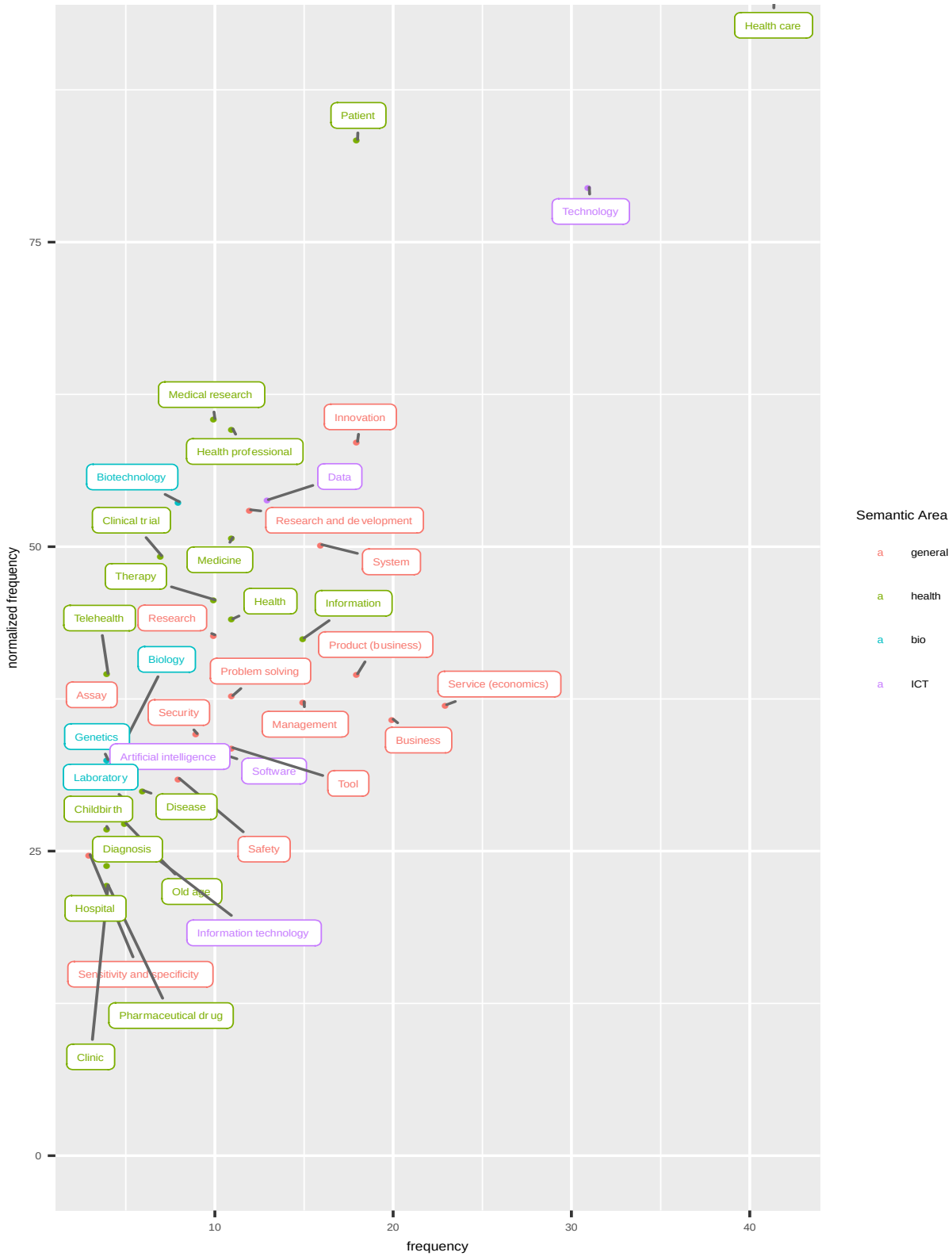
where $fr_{rs}(t)$ is the frequency of entity *t* in the restricted sample, $fr_{es}(t)$ is the frequency of the entity in the extended sample and *N* is the number of websites in the extended sample. The ratio in brackets gets smaller when an entity occurs frequently in the extended sample.

The entities plotted in Figure 2 can be split into four semantic areas. Three of these areas are quite specific: health, bioscience and ICT. The fourth area is rather general as it includes entities such as: *problem solving*, *management*, *business*, and *tool*. Although many of these general entities often appear in the restricted sample, they are not given high normalized scores due to their high frequency in the extended sample. Because the normalized score is more effective in reducing the incidence of general entities in the benchmark sample, we use this score to assign entities to the benchmark set.

The initial benchmark set is then expanded to include some additional entities recovered from the extended sample. Recovered entities are semantically close to the ones in the benchmark set, and their inclusion reduces the risk that some important sector-specific entities, which do not occur in the restricted sample, may be missed. To identify semantically close entities we use the Word-to-Vector model (W2V) (Mikolov et al., 2013).[11] The basic principle of this model is to map each entity, into a vector of coordinates that locates that entity in a multidimensional space where it is possible to compute geometrical distances from other entities.[12] By measuring the closeness between each entity in the benchmark set and any other entity from the extended sample, we can find new elements to augment the benchmark set. For instance, when searching for entities close to *mhealth* (included in the benchmark set) we find: *shared care*, *NHS health check*, *acute care*, *vital signs*. Each of these terms can be added to the initial group of terms so as to obtain an expanded benchmark set.

[11] In this report we somehow improperly refer to the Word-to-Vector model and to W2V scores because distances are computed based on entities rather than on individual words.
[12] A description of the Word-to-Vector model, together with references, can be found at the following address https://code.google.com/archive/p/word2vec/.

Figure 2: Digital Health benchmark entities (subsample)



Notes: The figure plots the most frequent entities extracted from the websites of companies in the restricted sample. Entities are plotted according to their simple frequency (number of websites from which the entity is extracted) on the x-axis, and their normalized frequency (correcting for overall frequency) on the y-axis.

Table 4: Ten highest scoring entities for each sector

Entity	$nfr(t)$
Digital health	
Health care	144.2
Patient	83.4
Technology	79.5
Telecare	76.7
Medical research	60.5
Health professional	59.7
Innovation	58.6
Data	53.9
Biotechnology	53.7
Research and development	53
Financial sector	
Insurance	9246.8
Investment	6116.2
Pension	5546.4
Finance	4619.8
Financial adviser	4423.1
Mortgage loan	4199.2
Financial Conduct Authority	3167.8
Independent Financial Adviser	3163.5
Broker	3142.2
Market	3136.5
Processing industry	
Chemical industry	36.0
Raw material	34.0
Chemical substance	29.9
Organic chemistry	29.6

Entity	$nfr(t)$
Pharmaceutical industry	28.5
Chemistry	26.7
Fine chemical	26.3
Chemical engineering	24.1
Pharmaceutical drug	22.2
Polymer	20.3

Notes: For each of the three sectors the table provides the list of the 10 entities with the highest normalized frequency score.

For each sector, Table 4 lists the 10 entities with the highest normalized frequency score. Scores should not be compared across sectors because their magnitude varies with the size of the restricted sample. Because the restricted sample for the Financial Sector is much larger than the ones for the other two sectors, normalized frequencies are much greater.

2.3 Company classification

We use the sector-specific ***extended benchmark sets of entities*** to identify, within the extended sample, companies that may belong to the sector even if they are excluded from the restricted sample. The W2V model is now used to compute a measure of "closeness" between a company's entities and the entities in a sector's *expanded benchmark set*. Companies with closer entities are assigned higher W2V score.

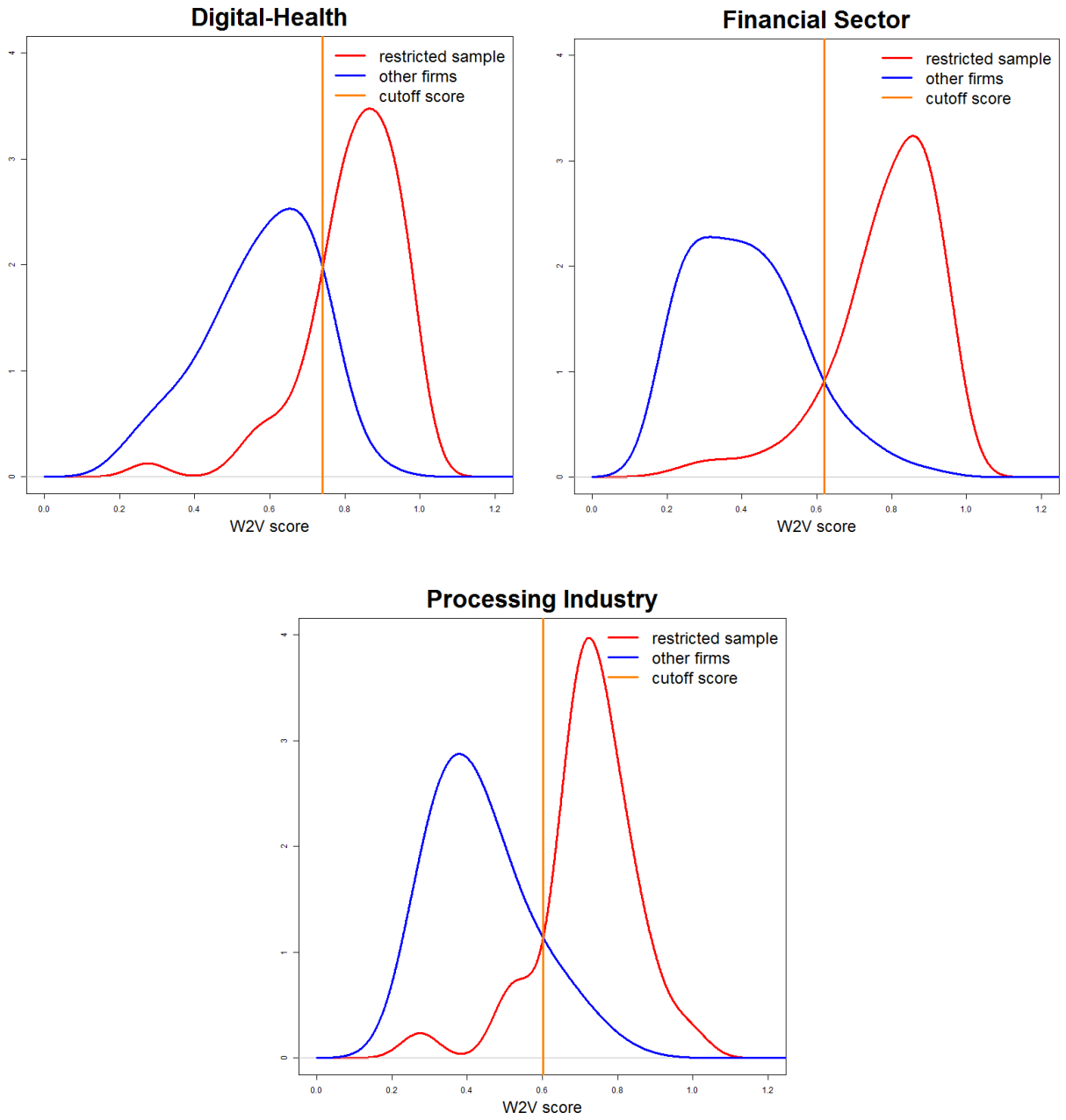
For the Digital-Health sector, we complement the W2V score with a second score called Elastic Query (EQ). The joint use of the W2V and the EQ scores improve the classification performance in the Digital-Health sector while it does not generate performance gains for the other two sectors.[13] The EQ score is obtained by jointly performing a search on the websites of the extended sample. This query searches in the text of the companies' websites all terms from the *extended benchmark set.* The EQ score reflects a company's ranking in the list of query results. A company gets a high Elastic Query score if its website contains many of the entities in the *extended benchmark set.*

Our first classification task is to identify companies that have a "sufficiently high probability" of being part of the sector within the extended sample. To achieve this goal, we need to set a W2V score cut-off. Companies with a W2V score below the cut-off are classified as "not part of the sector". To find this cut-off we first estimate the probability distribution of the scores across two groups of companies: those excluded and those included in the restricted sample. In other words, for each level of the W2V score we obtain separate probabilities of observing that score within the restricted sample and within the sample of all the other firms. We then set the W2V cut-off at the intersection between the two

[13] Due to the limited timeframe of the study, classification performance is evaluated by checking manually a limited sample of randomly drawn websites from companies classified with different scores. We acknowledge that a more systematic and time-consuming approach would be useful in order to evaluate and to improve the performance of the classification exercise.

probability distributions. Scores above this level have a higher probability of occurring in the restricted sample than to occur in the sample of all the other companies. W2V cut-offs for individual sectors are illustrated in Figure 3.

Figure 3: Identification of the W2V cut-offs for each sector



Notes: The figure shows the empirical distribution of the W2V scores. For each sector we separately estimate the distribution within the restricted sample and within the group of other firms in the extended sample. The W2V cut-offs are set at the intersection of these distributions.

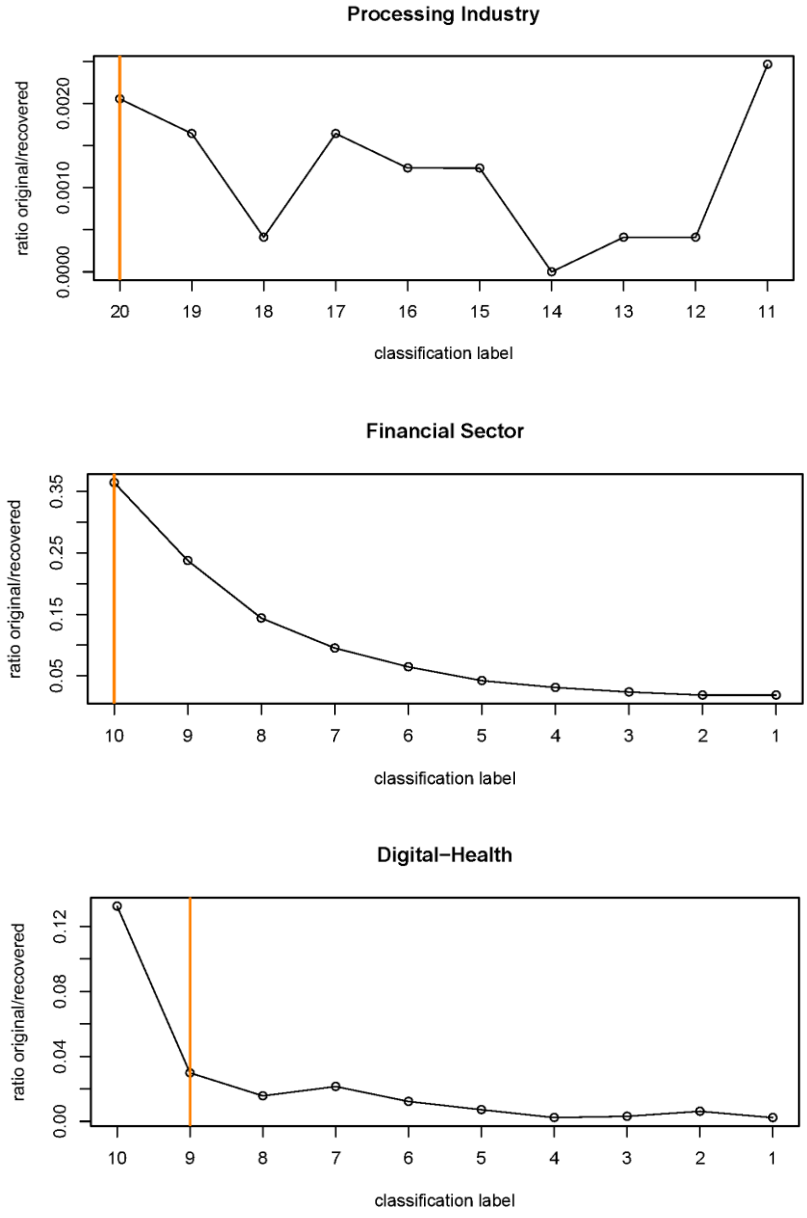
Admittedly, this classification strategy is far from perfect. First, the restricted samples for the Digital Health sector and the Processing Industry are rather small generating imprecise estimates of the W2V distribution within these samples. Second, an optimal classification strategy would require comparing the W2V distribution within the restricted sample and within a sample of companies that includes only companies that do not belong to the

sector. Unfortunately, this can be achieved only by checking manually a number of random samples to exclude those that we recognize (based on some prior knowledge) as part of the sector. Because this process is extremely time-consuming, given the limited time-frame of this project, we implement a second-best classification strategy.

Companies with a W2V score above the cut-off are defined as “recovered” companies. The next step is to split recovered companies in groups with different probabilities of being part of the sector. We do so by assigning to each company a classification label based on the quantile of its W2V score within the distribution of this score among recovered companies only. Recovered companies with a higher value of the classification label are more likely to belong to the sector as their entities are “closer” to the ones of the benchmark set.

For each sector, we provide to BIS a dataset listing all recovered companies and their classification label. To minimize the incidence of false-positives when we investigate geographical clustering, we use for this analysis only companies with the top classification labels. For the Digital-Health sector we retain recovered companies with value labels 9 and 10 (i.e., whose w2v score is respectively above the 8^{th} and the 9^{th} deciles of the w2v distribution among recovered companies).

Figure 4: Ratio of companies in the restricted sample over recovered companies by classification label



Notes: The figures show the proportion of original over recovered companies by different classification labels. The vertical line indicates the level of the classification label below which recovered companies are excluded from cluster analysis. For the processing industry, we split the sample in 20 quantiles instead of 10 quantiles. The reason for doing this is that the number of recovered companies is much larger for this sector.

Figure 4 shows the proportion of restricted sample companies for each classification label of the w2v score. A possible approach to determine the label (i.e., the minimum classification label for inclusion in the cluster analysis) is to identify the label at which the proportion of restricted sample companies drops sharply. We can identify a discontinuity that fits this approach in the case of Digital-Health. However, this approach would not be appropriate for the Financial Sector where the proportion decreases gradually, of for the Processing Industry where the proportion does not decrease monotonically as we move from higher to lower labels. We eventually choose to set the cutoff labels manually: 10 and

9 for Digital-Health, 10 for the Financial Sector and 20 for the Processing Industry (i.e., here we divide the distribution in 20 ventiles instead of 10 deciles). These labels where selected after conducting manual checks to verify the incidence of false positives within the classification labels retained for analysis.

Checks on the websites of classified Digital-Health companies suggest that the incidence of "false positives" is a serious issue within specific SIC classes. In particular, we tend to identify many "false" Digital-Health companies among wholesalers selling health products, and residential care companies.[14] Therefore, we decided to restrict the identification of Digital-Health companies to specific industries: manufacturing (SIC: 10–33), Information and Communication Activities (SIC: 58-63), Professional Activities (SIC: 69–75), Rental (SIC: 77), Human Health Activities (SIC: 86), Non-residential care (SIC: 88). Table 5 reports the number of recovered companies across different classification labels; we highlight the cells reporting the number of companies retained for cluster analysis.

Table 5: Number of recovered firms by classification label

Digital Health		Financial Sector		Processing Industry	
Label	**Recovered**	**Label**	**Recovered**	**Label**	**Recovered**
10	216	**10**	2,517	**20**	2,428
9	325	**9**	3,019	**19**	2,430
8	376	**8**	3,389	**18**	2,432
7	456	**7**	3,584	**17**	2,430
6	564	**6**	3,704	**16**	2,430
5	691	**5**	3,793	**15**	2,431
4	861	**4**	3,837	**14**	2,433
3	961	**3**	3,866	**13**	2,433
2	1,124	**2**	3,886	**12**	2,432
1	1,314	**1**	3,886	**11**	2,428

Notes: The table reports the number of recovered companies by classification label. For the Digital Health and for the Financial Sector, we split the sample of recovered companies in 10 deciles corresponding to classification labels 1 to 10. For the Processing Industry, we split the sample in 20 ventiles (top 10 ventiles reported). We highlight the cells reporting the number of companies retained for cluster analysis.

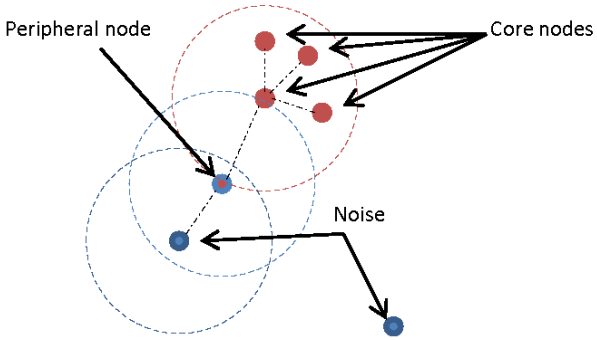
[14] After inspecting a number of websites of companies with these SIC codes we are inclined to conclude that most of the companies classified as Digital-Health do not employ (or sell) advanced technologies.

2.4 Identification of geographical clusters

The Code-Point[15] dataset is used to link the postcodes of the original and recovered companies to geographical coordinates (i.e., Easting and Northing). By using these coordinates we can locate each company in the UK territory represented as a continuous geographical space, making it possible to conduct an analysis of geographical clustering that does not require assigning companies to discrete geographical units.

We use the Density Based Spatial Clustering Algorithm with Noise (DBSCAN) introduced by Ester et al. (1996) to analyse spatial clustering of firms from the same sector in UK territory. This algorithm performs particularly well in detecting spatial concentrations of units with arbitrary shape within large databases. An advantage of this tool vis-à-vis other clustering algorithms is that it does not require prior knowledge of the number and of the location of clusters. Because DBSCAN does not require domain specific knowledge (e.g., specific knowledge of a particular industry) this algorithm can be easily applied to the study of very different sectors.

Figure 5: DBSCAN cluster identification



Notes: The diagram exemplifies the identification of spatial clusters by DBSCAN.

The DBSCAN algorithm identifies a cluster as the collection of:

a) ***Central nodes***. Central nodes are defined as companies that have **at least** *n* other companies belonging to the same sector within a radius of *eps* Km.

b) ***Peripheral nodes***. Peripheral nodes are companies that have **fewer** than *n* companies belonging to the same sector within a radius of *eps* Km (i.e., so they do not qualify as central nodes) **BUT** are within the radius of at least one central node.

The diagram in Figure 5 exemplifies a cluster identified by the DBSCAN algorithm where *n* = 3 and *eps* is the radius of all the circles around each node. After setting the parameters the DBSCAN identifies as a cluster the set of the four Core Nodes (i.e., each of them has

[15] Contains Ordnance Survey data © Crown copyright and database right 2015. Contains Royal Mail data © Royal Mail copyright and database right 2015. Contains National Statistics data © Crown copyright and database right 2015.

at least *n* = 3 other nodes within the *eps* radius) plus the Peripheral node that falls within the radius of one of the core nodes. Notice that this Peripheral Node does not qualify as Core because it has only two other nodes within radius. The Noise nodes (i.e., that do not qualify as Core Nodes nor as Peripheral Nodes) are excluded from the cluster.

When using DBSCAN to identify geographical agglomerations of companies from the same industry we are still required to set the parameters *n* and *eps.* These parameters conceptually correspond to the minimum *density* and the maximum *spread* of the clusters we are trying to identify. Inevitably, by changing these parameters DBSCAN reveals a different number of clusters, including a different number of companies. Therefore, for each sector we run the DBSCAN algorithm by setting different values for the parameters *n* and *eps*. A comparison of the clusters resulting from different iterations of the algorithm provides information on the differences between different clusters in terms of concentration and spread.

While the DBSCAN algorithm is useful to identify areas where companies from particular sectors concentrate, it does not control for the overall concentration of economic activities within a geographical area. In other words, DBSCAN cannot reveal if companies from a particular sector are attracted to an area by the same forces attracting companies from all the other sectors, or instead if the concentration is the result of sector-specific pull factors. To address this limitation we complement the DBSCAN analysis by estimating a modified version of the algorithm where core nodes CN_s for sector *s* are defined as follows:

$$CN_s = \begin{cases} \boldsymbol{yes}\ if\ \dfrac{N_s[dist < eps]}{N[dist < eps]} > \dfrac{N_s}{N} \\ \boldsymbol{no}\ otherwise \end{cases}$$

where $\frac{N_s[dist<eps]}{N[dist<eps]}$ is the ratio of firms from sector *s* over firms from all sectors within the *eps* radius of the candidate node, $\frac{N_s}{N}$ is the global ratio (i.e., at the dataset-level) of firms from sector *s* over all firms. Intuitively, we now impose a more restrictive condition for core nodes. It is no longer sufficient to have at least *n* companies from the same sector within the *eps* radius, but it is necessary that the concentration of companies from sector *s vis-à-vis* the concentration of all the other companies within that radius is greater than the global average. In successive iterations of the algorithm we will also require the density to be greater than two times the national average. We expect this algorithm to be more appropriate to highlight the areas that have a comparative advantage in attracting firms from the sectors under analysis.

2.5 Link analysis

The web-links extracted from companies' websites can be used to trace a network of the functional relationships between companies and other institutions. This network can be conveniently represented in a dataset format as an edgelist: a table where each row corresponds to an "edge" linking two "vertices".[16] The first column of the table includes an identification code for the first vertex, namely the company generating the link (i.e., the company whose website reports the link). The second column contains information on the

[16] We provide the edgelist to BIS as one of the outputs from the projects.

second vertex, namely the "target" website. We classify three types of targets based on their web domain:

- **Government.** These are webpages of public bodies that include the string ".gov" in their domain. For example, this type of targets includes Local Enterprise Partnerships (e.g., Coast-to-Capital LEP: www.coast2capital.org.uk), central government websites (e.g., www.gov.uk) and websites of individual government departments.

- **Organisations.** In most of the cases these are webpages of not-for-profit organisations that include the string ".org" in their domain. For example, this type of targets includes regulators (e.g., the Financial Conduct Authority: www.fca.org.uk), business networks (e.g., British Chambers of Commerce: www.britishchambers.org.uk), charitable organisations (e.g., AgeUK: www.ageuk.org.uk).

- **Higher education institutions.** In most of the cases these are websites of universities and research centres that are identified by the sting ".ac.uk" in their domain.

- **Companies.** All remaining targets are classified as companies. We distinguish between companies that are classified as part of the sector under analysis and companies that are not.

For each sector we describe the main feature of the network graph and we trace patterns of link exchange between companies classified under different SIC codes. Unfortunately, within each sector we manage to acquire links from the websites of a relatively small proportion of firms. Because the number of links for a unique cluster is often small, a comparison of networks across clusters may be misleading. For this reason we prefer to conduct the analysis at the industry level.

While extracting websites from companies' website we tried to identify the functional relationship they represent (i.e., supplier, client, partnership) based on the headings of the webpage from which the link was extracted. Unfortunately, the number of links that we manage to classify is too small to be used as an input for analysis.[17]

[17] In the data file on links that we provide to BIS the variable "label" contains this link classification.

2. Quantitative analysis: results

In this section we present the results of the quantitative analysis. For each sector, we first describe the distribution of recovered and original companies across SIC codes, so as to evaluate the extent to which the data-drive identification of the sector departs from standard industry codes. We will then describe the patterns of geographical location identified by the DBSCAN algorithm and by the adjusted DBSCAN algorithm. Finally, we describe the network of web-links.

2.5 The Digital-Health sector

A recent report commissioned by the UK Office of Life Sciences defines Digital Health as "an emerging industry arising from the intersection of healthcare services, information technology and mobile technology" (Deloitte, 2015). This definition of the sector suggests that Standard Industry Classification (SIC) codes may be inappropriate to identify companies operating within this industry. The hybrid nature of this industry emerges also when grouping entities extracted from the websites of restricted sample companies. Table 6 splits the entities included in the extended benchmark set by semantic area identified through SpazioDati's natural language processing algorithm. It is clear that the entities extracted through website analysis are not associated with a unique SIC code, and that data-driven approaches are more appropriate to capture the hybrid nature of the industry. This evidence suggests that data-driven approaches to industry classification are particularly powerful in identifying industries that span across different economic activities that can be mapped to different semantic areas.

Table 6: Detailed division of Digital- Health entities by semantic areas

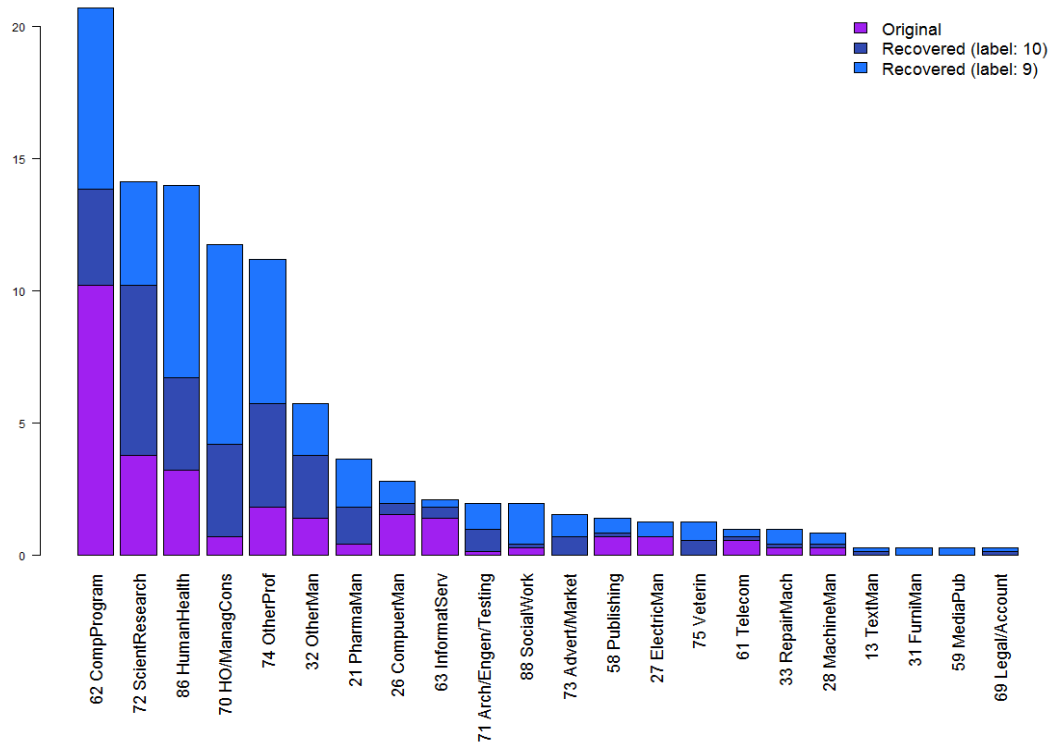
Segment	Entities
Telecare and telehealth i.e. support and assistance provided at a distance between clinics and patients. Patients' data are sometimes collected using mobile technologies and application.	Telecare Hospital Disability Health Medicine Molecular diagnostics Patient Quality of life Telehealth Childbirth Clinic Health informatics Therapy Summary Care Record Disease Vital signs Health professional MHealth Assisted living Health care Old age

Segment	Entities
Management/sales semantic area.	Information Research Regulation Management Risk Service (economics) Information technology Customer Empowerment Business Sales Problem solving Education
Health information storage and analytics using AI algorithms.	System Quality assurance Innovation Artificial intelligence Tool Research and development Data Software engineering Security Software Product (business) Computer Safety Technology
Biological, biomedical and genetic research. Pharmaceutical experimentation.	Genetics Genomics Microarray Gene expression Clinical trial Biology Medical research Assay Pharmaceutical drug Medical device Sensitivity and specificity Translational research Drug Polymerase chain reaction Biopharmaceutical Molecular biology DNA sequencing Diagnosis Laboratory Biotechnology Antibody Protein

Notes: The table lists all the entities included in the extended benchmark set for the Digital-Health sector. Entities are grouped by semantic area.

The fact that a unique SIC code cannot capture the full extent of the Digital-Health sector is apparent from Figure 6. The figure shows the percentages of original and recovered companies by different 2-digit SIC codes.[18] Figure 6 shows that most of the recovered companies fall within five SIC sectors: Computer Programming (SIC: 62), Scientific Research (SIC: 72), Human Health (SIC: 86), Head Office and Management Consulting Activities (SIC: 70), Other Professional Services (SIC: 74). While the first three SIC codes are very consistent with the definition of the sector as the intersection between technology and health activities, the last two codes suggest that some of the companies that are primarily engaged in consulting and broader professional activities are also related to Digital-Health. By inspecting the website of one of these consulting firms, we find that the company advertises on its website services provided to the health-care sector to promote the adoption of new technologies.

Figure 6: Percentages of original and recovered Digital-Health companies by SIC code



Notes: The figure shows the percentages (y-axis) of original and recovered firms for the Digital-Health sectors classified by 2-digit SIC code.

Table 7 lists the Digital-Health clusters identified by different iterations of the DBSCAN algorithm. Each cluster takes the name of the Travel-to-Work Area (TTWA) where most of its companies are based. This does not imply that we constrain a cluster to be contained within a unique TTWA. We estimate the DBSCAN algorithm with three values for the radius *eps*: 10km, 15km and 20km. The medium radius (15km) is our preferred setting as

18 Because we excluded the wholesale sector and the residential care sector from the website analysis (see section 2.3 Company classification) we exclude the corresponding SIC codes from the bar chart in Figure 6.

it is in line with the average commuting distance across the UK (ONS Census, 2011). For each value of the radius, we run DBSCAN with three different values for the density parameter *n* equal to 5, 10 and 15 companies. As we increase the density parameter from 5 to 15 the number of clusters identified by DBSCAN decreases because a greater number of more isolated firms falls into the category “Outside cluster”. Results do not change dramatically when we extend the radius from 10km to 15km. On the contrary, when we set the radius at 20km, we tend to identify few macro-clusters. For instance, when *eps*=20km the cluster of London absorbs both Cambridge and Oxford. In the table, we highlight the clusters that are more robust to changes in the parameters. DBSCAN with radius *eps*=15km and density *n*=10 identifies the most robust clusters. We map these clusters in Figure 7.

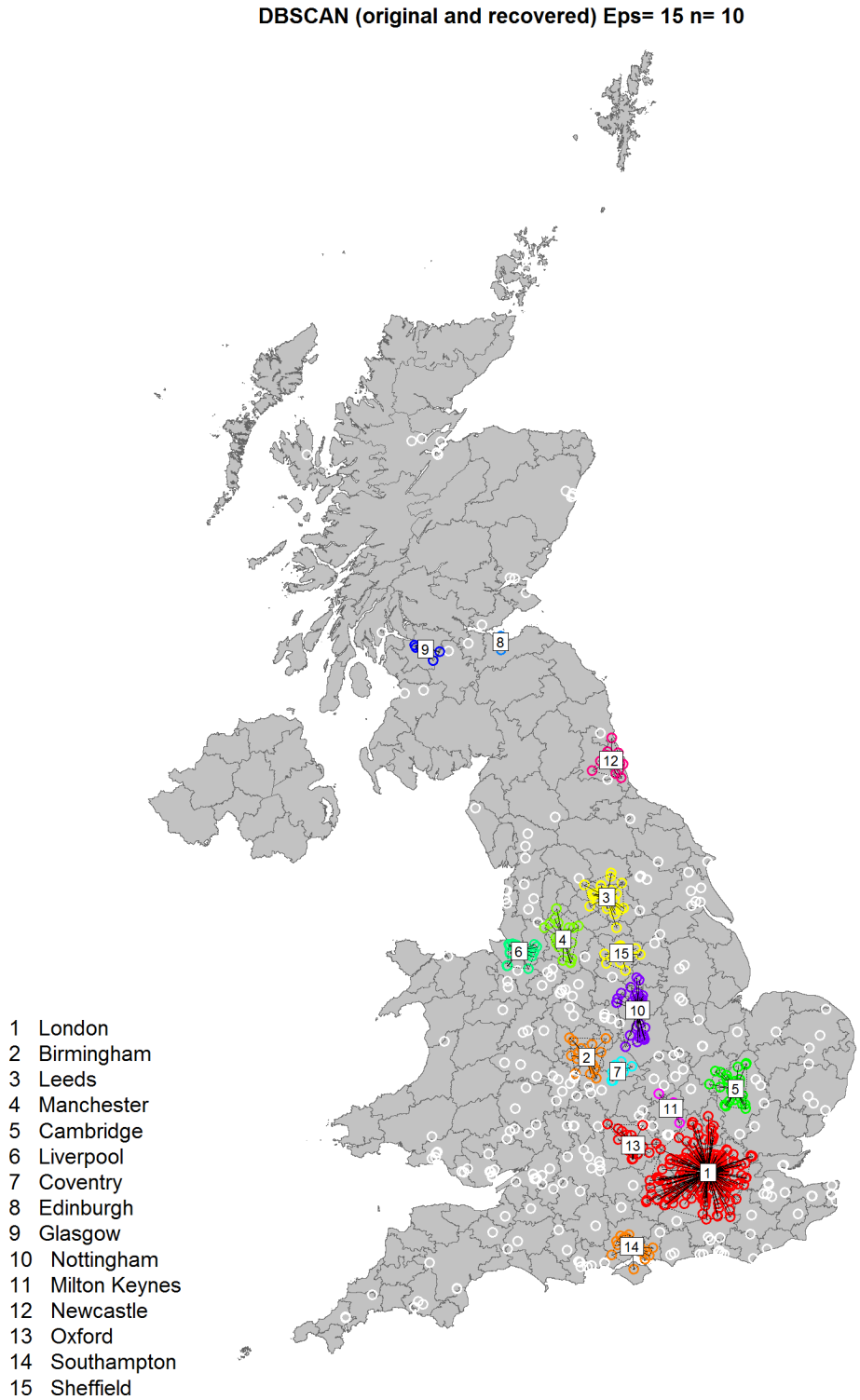
Table 7: Digital-Health clusters identified by DBSCAN (number of companies)

Main TTWA cluster	eps=10km			eps=15km			eps= 20km		
	n=5	n=10	n=15	n=5	n=10	n=15	n=5	n=10	n=15
Aberdeen	5			5			5		
Birmingham	23	22	22		29	29			48
Blackpool							5		
Brighton	5			11					
Bristol	8			9			12		
Burton upon Trent	9			9					
Cambridge	36	33	32	41	41	38			45
Cardiff	6			6			7		
Cinderford and Ross-on-Wye							5		
Coventry	10				10				
Edinburgh	13	13		13	13		13	13	
Glasgow	10	10		13	12		14	13	
High Wycombe and Aylesbury	5			5			6		
Leeds	27	24	15	33	32	31			36
Leicester	13	12							
Liverpool	12	12		26	23				22
London	276	232	207	349	298	288	425	411	350
Manchester	31	24	24	41	36	36	241	77	43
Margate and Ramsgate	5			5			7		
Medway	7			8					

Main TTWA cluster	eps=10km			eps=15km			eps= 20km		
Milton Keynes	9				11				
Newcastle	16	10		20	18	15	20		20
Nottingham	25	25	25	112	45	34		159	56
Oxford	28	28	16		31	30			
Poole	6								
Portsmouth	9								
Reading		20							
Sheffield	12	10			17	15			17
Southampton	12	12		33	23		35	28	23
Stevenage and Welwyn Garden City	14	10							
Stoke-on-Trent	5			10					
Swansea	5			5			6		
Swindon				6			7		
Warrington and Wigan	8								
Outside cluster	**241**	**394**	**550**	**131**	**252**	**375**	**83**	**160**	**231**
Total	**891**	**891**	**891**	**891**	**891**	**891**	**891**	**891**	**891**

Notes: For each cluster identified by the DBSCAN algorithm, the table reports the number of original and recovered (labels 10 and 9 only) companies. Cells are left empty when the cluster is not identified. Clusters take the name of the Travel-to-Work Area where most of the companies are based. We highlight the clusters that are more robust to different parameters of the DBSCAN algorithm.

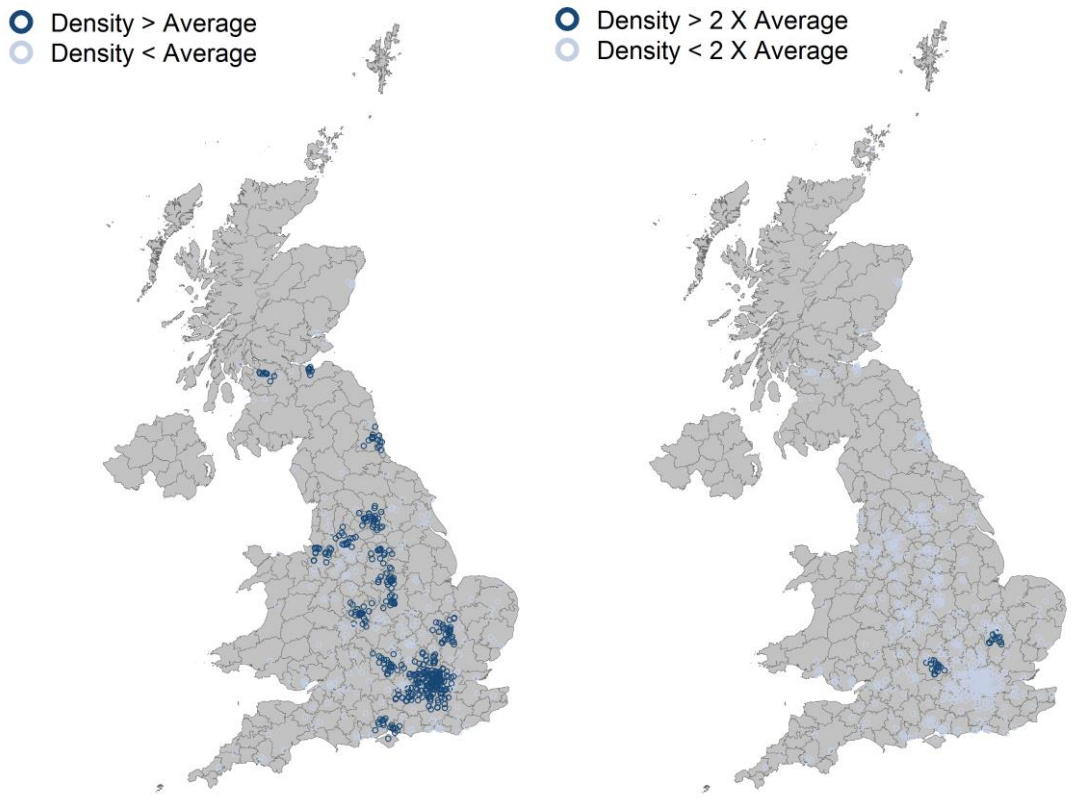
Figure 7: Map of Digital Health clusters



Notes: The figure shows the geographical location of Digital-Health clusters (colors) identified by running the BDSCAN algorithm with parameters eps =15km and n=5. White circles on the map represent companies that are not assigned to clusters.

The map in Figure 7 suggests that most of the clustering areas for the Digital-Health sector are close to London (i.e., Southampton, Oxford and Cambridge) or they locate along a South-North corridor stretching from Birmingham to Leeds. In the North, Edinburgh, Glasgow and Newcastle are identified as more isolated, and relatively smaller, clusters. The clusters identified by DBSCAN coincide to a large extent with the main urban areas. It remains to determine whether these clusters are still identified once we control for the overall concentration of companies belonging to other industries.

Figure 8: Digital Health clusters after controlling for other companies in clustering areas



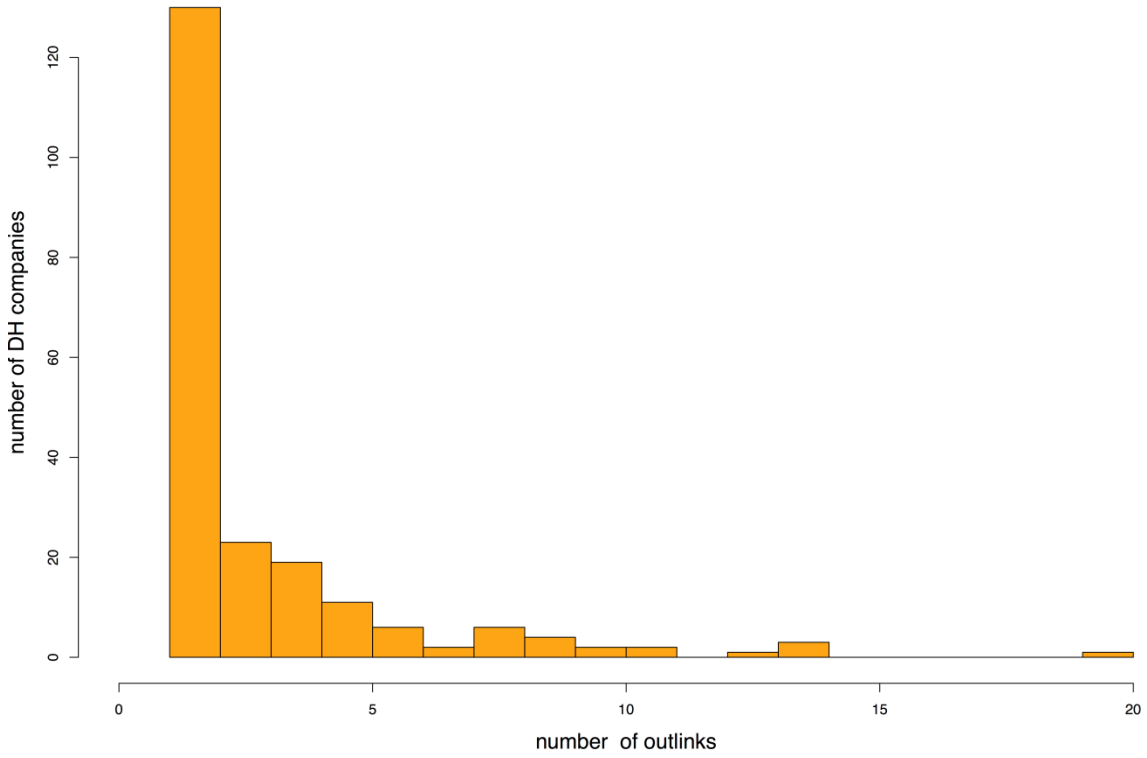
Notes: The two maps are obtained by running a version of the DBSCAN algorithm that controls for overall concentration of non-Digital-Health companies in the clustering area. Within the radius of companies plotted in darker blue there is a relative density of Digital-Health companies (obtained as the number of Digital Health companies in the radius over the total number of companies in the radius) greater than the national average (left-hand side panel), or greater than twice the national average (right-hand side panel).

Figure 8 shows the geographical areas where the density of Digital-Health companies is greater than the national average (left-hand side panel) or at least twice as great as the national average (right-hand side panel). Most of the clusters identified in Figure 7 are still shown in the left-hand side panel of Figure 8, and this suggests that there are sector-specific forces attracting Digital-Health companies to particular areas. Indeed, if urban areas were to attract companies from this sector as much as companies from any other sector, we would expect the number of Digital-Health companies to be higher in urban areas vis-à-vis non-urban areas, but their density to be the same as the national average. However, when we raise the bar of the test and require the density to be at least twice as the national average (right-hand side panel), we find that only Oxford and Cambridge are

still identified as clustering areas. This result suggests that institutional factors and local externalities are relatively stronger in these two areas compared to other clustering zones.

The outlinks extracted from the websites of Digital-Health companies are used to generate a network graph of the relationships between the following entities: companies, government institutions, higher education institutions and not-for-profit organisations. Out of 891 Digital-Health companies (original and recovered companies with classification label 10 and 9), we extract outlinks from the websites of 210 companies. Figure 9 shows the distribution of the 210 companies by the number of outlinks extracted from their websites. For a large number of these companies we extract only one outlink and the distribution is skewed toward low values (i.e., the maximum number of links extracted from a unique company's website is 20).

Figure 9: Distribution of Digital-Health companies by number of outlinks



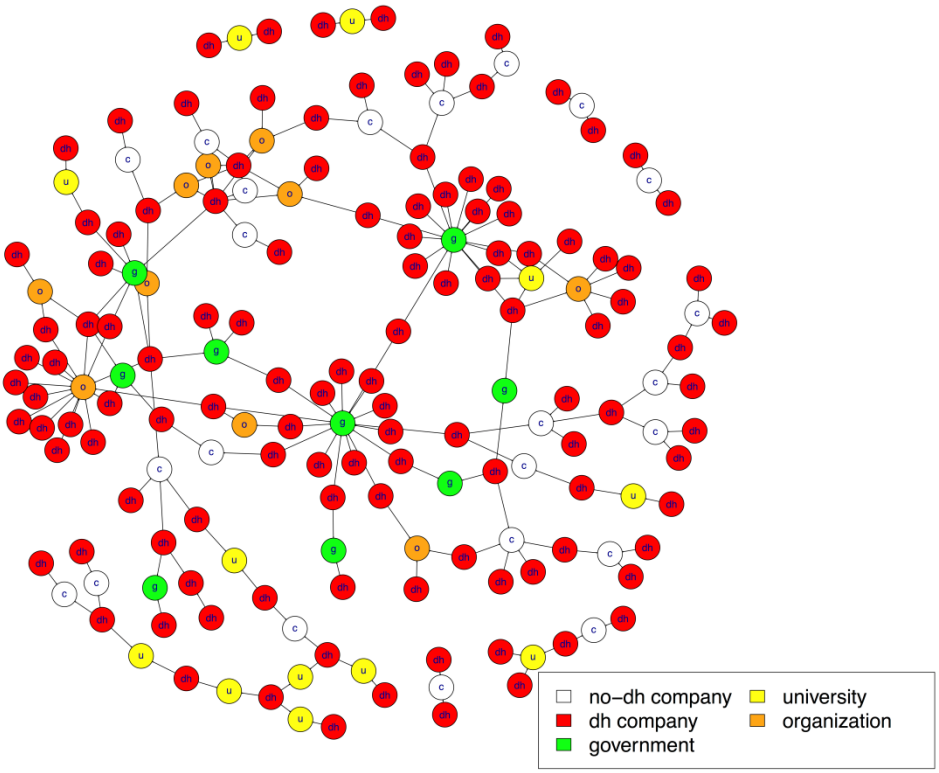
Notes: The figure shows the distribution of Digital-Health companies with outlinks by the number of outlinks extracted.

The relatively low number of links for each company may suggest that the Digital-Health sector includes many companies with a relatively small network of institutional or corporate partners. The limited extension of Digital-Health companies' networks is clearer in Figure 10. The figure plots Digital-Health companies from which the link is generated (i.e., red circles) and all target companies that attract at least two connections.[19] Edges between the vertices of the network graph represent one or more web-links. It is apparent that there are very few direct connections between Digital-Health companies. On the contrary, we

[19] This condition is necessary to avoid excessive cluttering of the graph.

observe many links from their websites pointing to the same websites of government institutions (i.e., green circles). Higher education institutions (i.e., yellow circles) and not-for-profit organisations have also a very central position in the network. Despite the fact that is difficult to infer the exact nature of the functional relationship expressed by a web-link connecting two websites, the relevance of institutional targets (i.e., government, universities and not-for profit) in the Digital-Health link network suggests the importance of corporate-institutional relationships for this sector.

Figure 10: Network graph of the web-links generated by Digital-Health companies



Notes: The network graph represents the links that from the websites of Digital-Health companies point to the websites of other companies (either Digital-Health or not) or institution. The graph represents only the links to target entities with at least two connections.

For each geographical cluster, Table 8 reports the number of links by type of target. Links to government websites constitute a high proportion of all the links across most of the clusters. Links to higher education websites are especially frequent across companies in London and Birmingham. Unfortunately, the small number of links for most of the clusters makes it difficult to draw robust conclusions on the qualitative differences of the network across geographical areas.

Table 8: Number of links by cluster and type of target (all sources are Digital-Health)

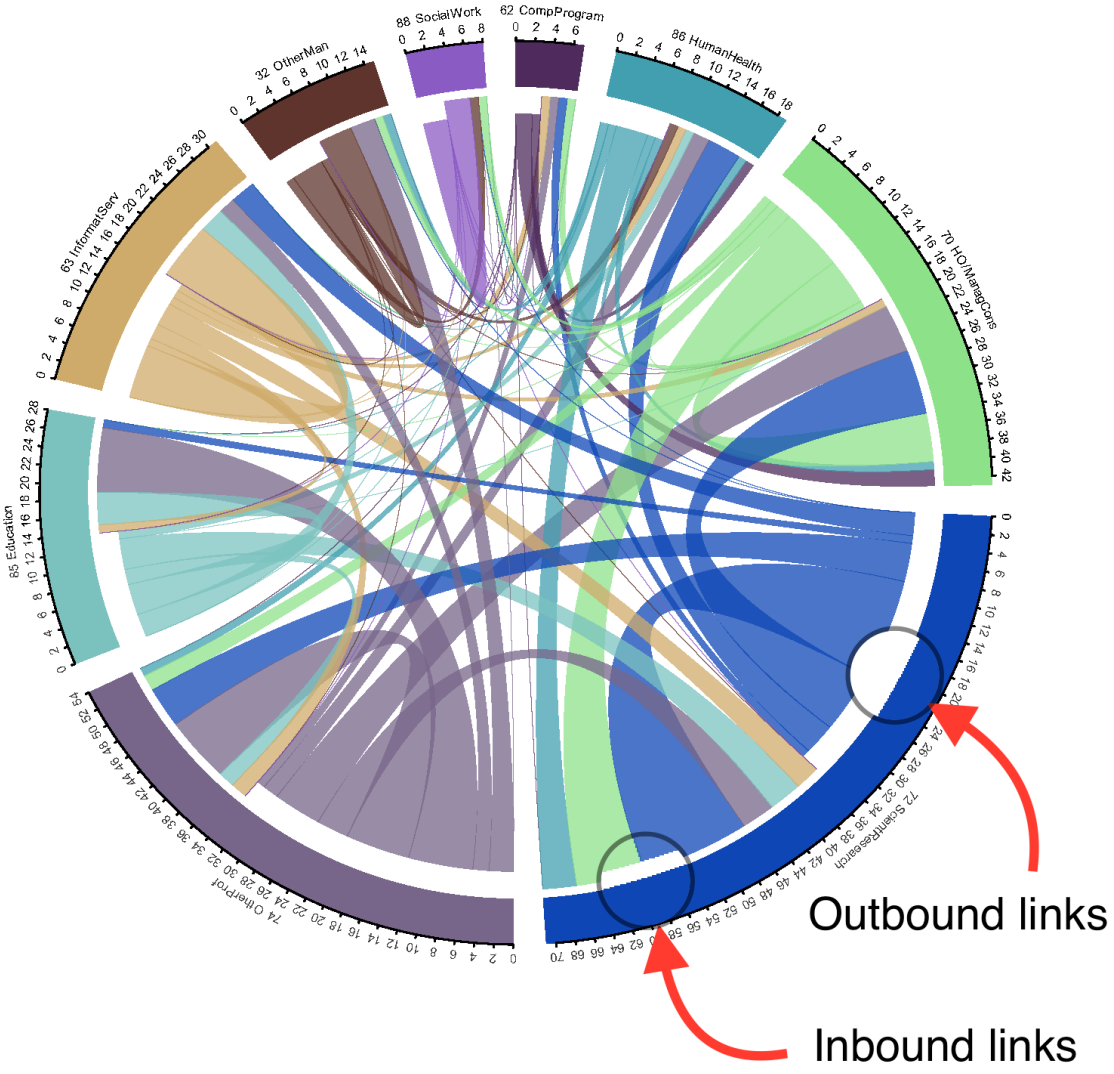
	company	government	organisation	university
London	104	24	24	20
Birmingham	120	42	30	21
Leeds	3	2	1	3
Manchester	9	1	7	1
Cambridge	7	3	7	3
Liverpool	13	2	8	6
Coventry	6	0	1	1
Edinburgh	5	0	0	0
Glasgow	8	6	0	7
Nottingham	11	3	0	7
Newcastle	6	4	4	7
Oxford	12	1	1	4
Southampton	14	12	4	3
Sheffield	3	1	1	0
Not in cluster	33	1	3	3

Notes: For each cluster, the table reports the number of links extracted from the webpages of Digital-Health companies by type of target.

The circle plot in Figure 11 shows the patterns of web-links across companies belonging to different SIC industries. As for the previous analysis, all links are generated by Digital-Health companies but they can be directed also to non-Digital-Health companies. The outer ring in the plot shows the number of links (both outbound and inbound) distributed across the nine most common SIC industries. The size of the flows crossing the circle is proportional to the number of links exchanged across SIC sectors. Inbound links (i.e., links pointing to companies within the sector) are represented closer to the outer ring, while outbound links (i.e., links generated by the companies within the sector) are more distant from the outer ring. For most of the SIC industries in the diagram, a large proportion of the links generated are directed to the companies belonging to the same industry. For example, eleven links generated by companies in SIC 72 (Scientific Research) are directed to other companies with the same SIC code. Nevertheless, there are also frequent connections across SIC classes. For instance, we identify eight links from the Management Consulting industry (SIC 70) to the Scientific Research industry (SIC 72) and numerous connections between Other Professional Services and other sectors. The limited overall number of links extracted from Digital-Health companies call for some caution in drawing conclusions on the functional relationships between industries.

Nevertheless, this analysis is suggestive of the fact that the Digital-Health sector promotes the integration of a wide range of economic activities.

Figure 11**: Web-links originated by Digital-Health companies across SIC industries**



Notes: The circle plot shows the flows of web-links between different SIC industries. All links are generated by companies classified as part of the Digital-Health sector but they are allowed to point to companies outside this sector. The outer ring represents the ten most frequent industries from (and to) which we observe web-links. Labels on the external side of the outer ring refer to the number of links. The size of the flows crossing the circle increases with the number of web-links exchanged. Links generated by companies from (to) one SIC class are more distant from (closer to) the section of the outer ring for that class.

2.5 The Financial Sector

The Financial Sector offers the opportunity to test the classification methodology on a population of companies that are perhaps better represented by SIC classes. A possible gain from exploiting website data is the possibility to identify companies that offer financial services while conducting other economic activities and that are classified under SIC codes not related to the financial industry. Consistently with the greater homogeneity of

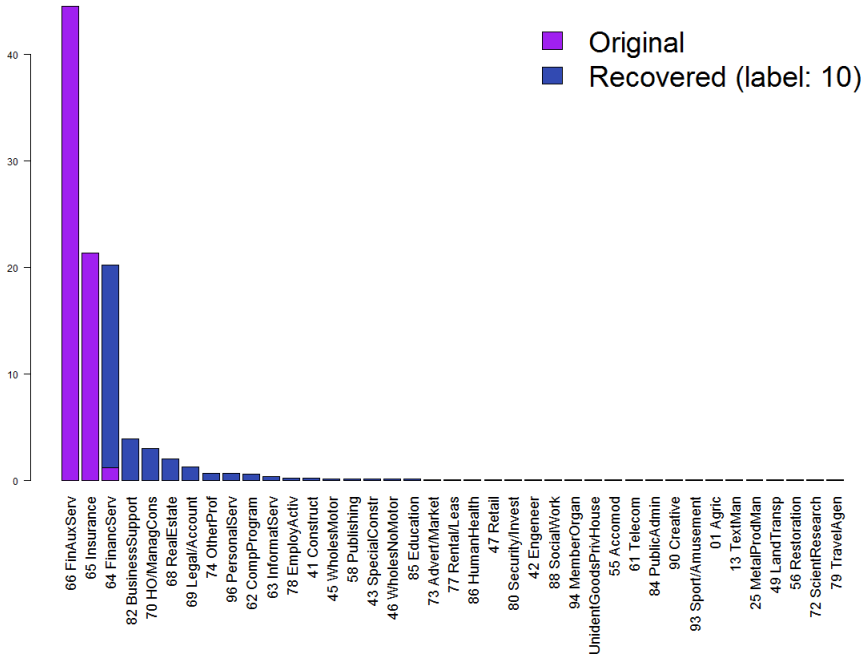
this sector vis-à-vis the Digital-Health sector, most of the entities extracted from the websites of restricted sample of companies can be assigned to the semantic area of finance and accounting (see Table 9).

Table 9: Top 80 entities (by normalized score) for the Financial Sector

Insurance	Investment management	Trust law	Property insurance
Investment	Capital (economics)	Bank	Mergers and acquisitions
Pension	Financial planner	Option (finance)	Funding
Finance	Vehicle insurance	Trade	Contents insurance
Financial adviser	Insurance policy	Tax	Rate of return
Mortgage loan	Loan	Reinsurance	Financial market
Financial Conduct Authority	Financial plan	Legal liability	Institutional investor
Independent Financial Adviser	Saving	Short (finance)	Private equity
Broker	Risk management	Brokerage firm	Equity (finance)
Market (economics)	Money	Retirement planning	Mortgage broker
Risk	Wealth	Debt	Real estate
Financial services	Investment fund	Economic growth	Interest rate
Life insurance	Health insurance	Legal personality	General insurance
Insurance broker	Retirement	Stock	Self-invested personal pension
Underwriting	Asset	Liability insurance	Defined benefit pension plan
Home insurance	Travel insurance	Corporate finance	Bond (finance)
Income	Portfolio (finance)	Sales	Deposit account
Wealth management	Asset management	Financial Services Authority	Equity release
Property	Accounting	Buy to let	Commercial bank
Security (finance)	Investor	Consultant	Professional liability insurance

Notes: The table lists the 80 most frequent entities extracted from the restricted sample of the Financial Sector.

Figure 12: Percentages of original and recovered Financial Sector companies by SIC code



Notes: The figure shows the percentages (y-axis) of original and recovered firms for the Financial Sectors by 2-digit SIC code.

Table 10: Finance clusters identified by DBSCAN (number of companies)

	eps=5km			eps=10km			eps= 15km		
	n=20	n=30	n=50	n=20	n=30	n=50	n=20	n=30	n=50
Bath	22								
Birmingham	80	73	58	325	264	178		692	329
Blackburn					37				
Bournemouth	89	68		104	103	103		113	113
Brighton	47	45			77	69			
Bristol	64	60		147	136	131		186	148
Cambridge				20					
Canterbury				25					
Cardiff	66	64		94	84	83	154	108	98
Chelmsford	29								
Cheltenham	34			62	57				67
Chester	27								
Colchester	27			44	41				
Coventry	20								
Derby	31								
Dudley	40	28							
Eastbourne	20			25				34	
Edinburgh	86	84	84	95	95	95	106	106	106
Exeter	21			29			61	35	
Glasgow	80	71	70	114	112	109	122	119	118
Guilford	53	33							
Harrogate	23								
Huddersfield	20								
Ipswich				28					
Leamington Spa	23								
Leeds	107	65	54	224	201	160			240
Leicester	58	47		67	67	62			90
Liverpool	50	48		116	118	86			
London	1948	1822	1691	2780	2613	2183	5219	3079	2833

	eps=5km			eps=10km			eps= 15km		
Luton	37								
Manchester / Liverpool	234	192	155	497	417	341		1015	643
Medway	35	30				54			
Middlesbrough and Stockton				26				45	
Milton Keynes				27					
Newcastle	43	41		92	85	74	163	108	107
Northampton	35			45	47				52
Norwich	42	39		52	52		58	58	55
Nottingham	62	56		120	112	75			142
Oxford				30				31	
Peterborough				32			39	34	
Plymouth	20			28			32		
Preston	34								
Reading	28					46			
Sheffield				51	42				
Shrewsbury							31		
Southampton	33			108	106	69	232	119	117
Southend	58	58	50			68			
Stoke-on-Trent	27			45	40				
Swansea				28				36	
Swindon				22			23		
Taunton				21			29		
Tunbridge Wells	49					32			
Outside cluster	**2989**	**3867**	**4629**	**1268**	**1885**	**2773**	**522**	**873**	**1533**
Total	**6791**	**6791**	**6791**	**6791**	**6791**	**6791**	**6791**	**6791**	**6791**

Notes: For each of the clusters identified by the DBSCAN algorithm, the table reports the number of original and recovered (only label=10) companies. Cells are left empty when the cluster is not identified. Clusters take the name of the Travel to Work Area where most of the companies are based. In the table, we highlight the clusters that are more robust to different parameters of the DBSCAN algorithm.

Figure 12 shows that the greatest majority of recovered firms fall within a few SIC classes. The restricted sample now includes all companies with SIC codes 65 (i.e., banking and auxiliary financial services) and 66 (i.e., insurance). This is the reason why the first two columns of the bar chart represent only original companies. The concentration of recovered companies in few SIC classes suggests that the classification methodology

based on website data provides little additional information on companies' activity compared to traditional industry classes.

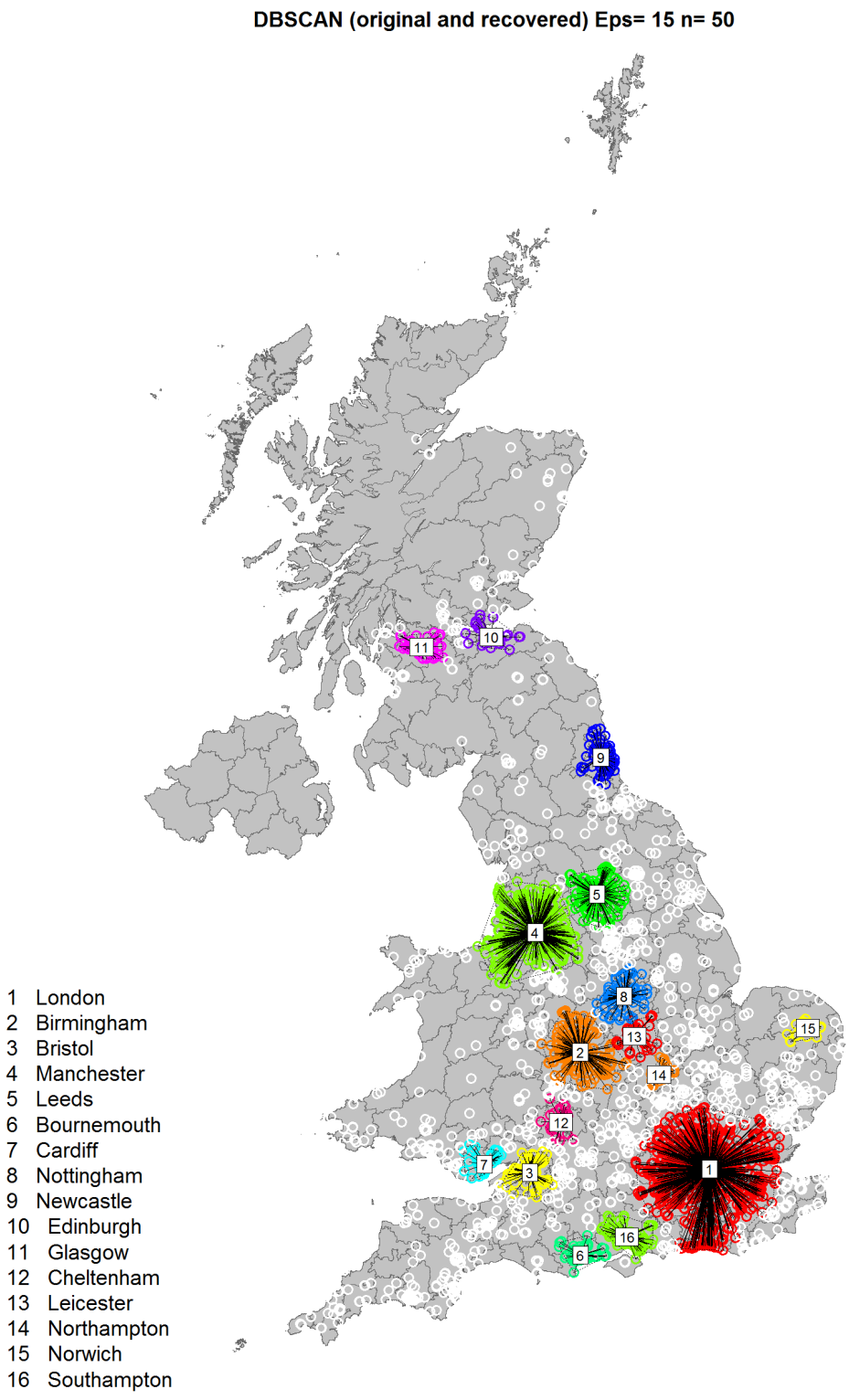
When iterating the DBSCAN algorithm for the identification of the geographical clusters, more restrictive density and radius parameters are set to account for the larger number of companies in this sector. In successive iterations, the density parameter *n* takes values 20, 30 and 50, and the radius *eps* takes values 5km, 10km and 15km. Table 10 lists the clusters identified during different iterations of the algorithm and the number of companies assigned to each cluster.[20] For values of *eps* below 15km and values of *n* below 50, DBSCAN identifies a large number of very small clusters, often splitting large urban areas in multiple clusters. For this reason, the values *n* = 50 and *eps* = 15km are chosen as the baseline parameters of the algorithm. The resulting clusters are mapped in Figure 13.

As expected, the cluster of London dominates the sector with over 2,800 companies identified. Second by size, with 643 companies, is the cluster that includes the metropolitan areas of Manchester and Liverpool. Leeds is a smaller cluster that is identified in most of the DBSCAN iterations. Surprisingly this cluster is absorbed by the Manchester/Liverpool cluster when *eps* = 15km and *n* is lower than 50. This happens because for lower values of the density parameter the algorithm captures a "corridor" of more isolated financial companies merging the three urban areas. In the North, Edinburgh, Glasgow and Newcastle emerge as smaller clusters. Birmingham, Nottingham and Leicester form a continuum of three clusters in the Midlands, while Bristol and Cardiff emerge as financial clusters in the South West.

Figure 14 shows the clustering areas where the ratio of financial companies over all companies is greater than the national average (left-hand side panel), and where this ratio is higher than twice the national average (right-hand side panel). While reduced in geographical extension compared to the ones identified in Figure 13, within most of the financial sector clusters previously identified there are areas of high "relative density" greater than the national average. On the contrary, no area has relative density of financial companies that is at least twice the national average. This evidence contrasts with previous findings for the Digital-Health Sectors where Oxford and Cambridge were identified as areas of high sectoral density. A possible explanation for these diverging results is that for the financial sector we are capturing a much larger and heterogeneous population of companies that is attracted to urban areas by factors that are common to many companies from other sectors (e.g., access to a larger pool of clients).

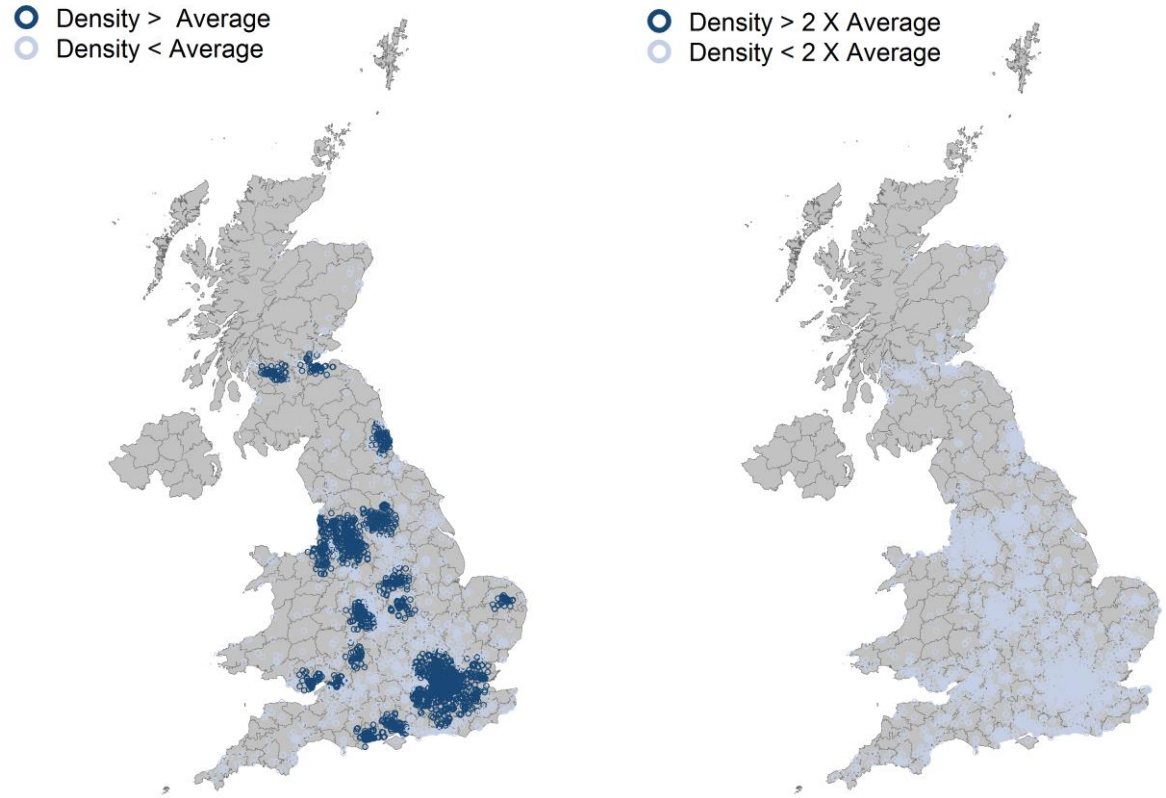
[20] Each cluster is named after the name of the TTWA where most of its companies are based.

Figure 13: Map of Financial Sector clusters



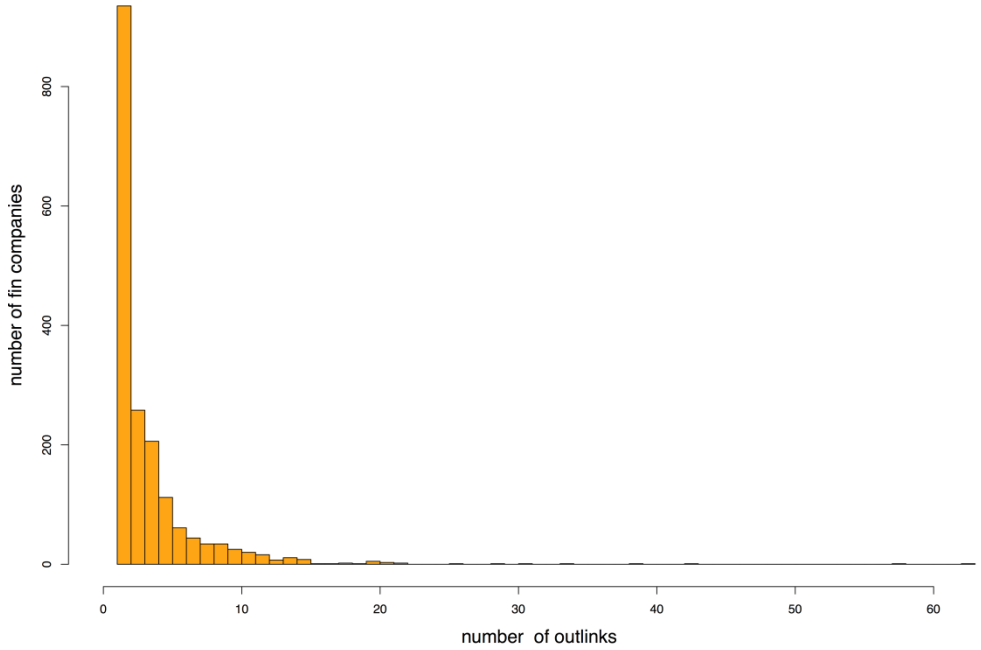
Notes: The figure shows the geographical location of the Financial Sector clusters (colours) identified by running the DBSCAN algorithm (eps =15km, n=50). The map is obtained by locating geographically all original and recovered companies with label=10. White circles on the map represent companies that are not assigned to clusters.

Figure 14: Financial clusters after controlling for non-financial companies in clustering areas



Notes: The two maps are obtained by running a version of the DBSCAN algorithm that accounts for overall concentration of non-financial companies in the clustering area. Within the radius of companies plotted in darker blue there is a relative density of financial companies (obtained as the number of financial companies in the radius over the total number of companies in the radius) greater than the national average (left-hand side panel), or greater than twice the national average (right-hand side panel).

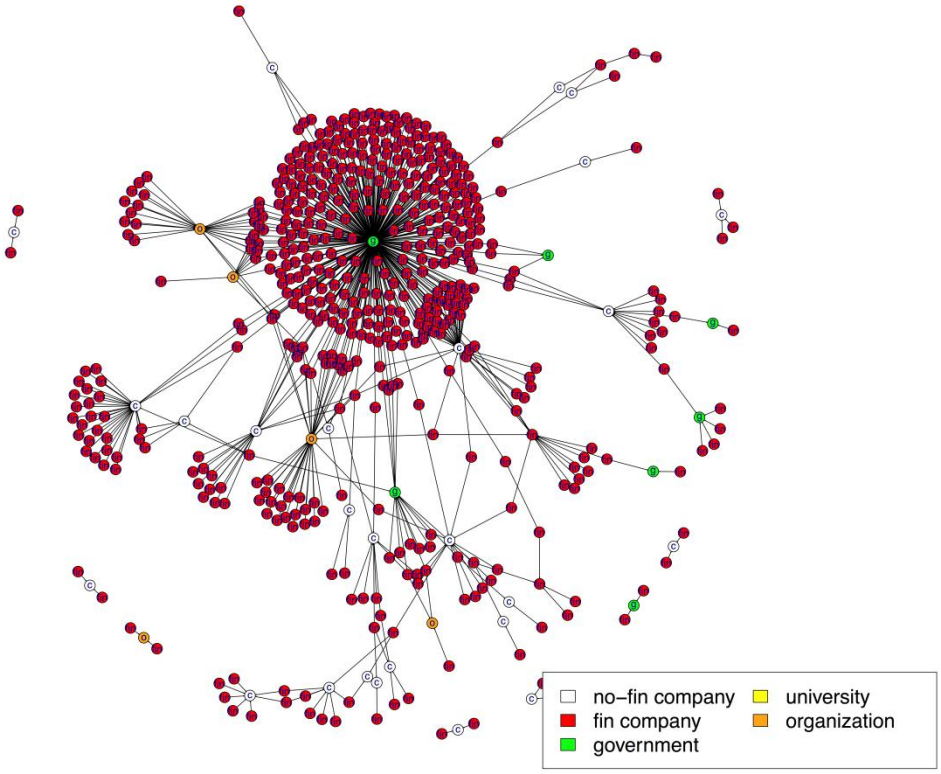
Figure 15: Distribution of financial companies by number of outlinks



Notes: The figure shows the distribution of financial companies by the number of outlinks extracted.

The distribution of financial companies by number of outlinks extracted is similar to the one for the Digital-Health sector: except for a small number of firms at the top end of the distribution, most companies have less than ten outlinks (Figure 15). In contrast, some interesting differences emerge when comparing the network graph of weblinks for the Financial Sector in Figure 16 with the same graph for the Digital-Health sector in Figure 15. First, no-link to higher education institutions can be identified in the network graph for the Financial Sector. This evidence contrasts with the important role played by these types of vertices in the network graph for the Digital-Health sector. For the Financial Sector we find a very large number of companies that include on their websites links to government websites (e.g., in Figure 16, the "government" website that attracts a large number of connections from financial companies is the web-domain www.gov.uk) or to the website of regulatory agencies (e.g., the Financial Conduct Authority website www.fca.org.uk). After a series of checks on individual companies' websites we conclude that these links are often included to redirect perspective clients to information on regulations and professional standards. Web-links connecting directly the websites of financial companies are also more frequent than those connecting digital-health companies.

Figure 16: Network graph of the web-links originated by Financial Sector companies



Notes: The network graph represents the links that from the websites of financial companies redirect to the websites of other companies (either from the same sector or not) or institutions. To avoid excessive cluttering of the graph we represent a 10% randomly drawn sample of target entities.

Table 11: Number of links by cluster and type of target (all sources are Financial)

	company	government	organisation	university
London	1687	558	286	42
Birmingham	206	92	39	2
Bristol	73	31	12	0
Manchester	350	164	61	0
Leeds	166	75	33	0
Bournemouth	68	43	9	2
Cardiff	40	32	16	0
Nottingham	87	40	18	0
Newcastle	59	32	9	1
Edinburgh	99	25	15	4
Glasgow	62	28	8	0
Cheltenham	35	14	5	0
Leicester	72	23	10	0
Northampton	19	10	3	0
Norwich	30	14	8	0
Southampton	47	34	22	4
Not in cluster	830	412	174	20

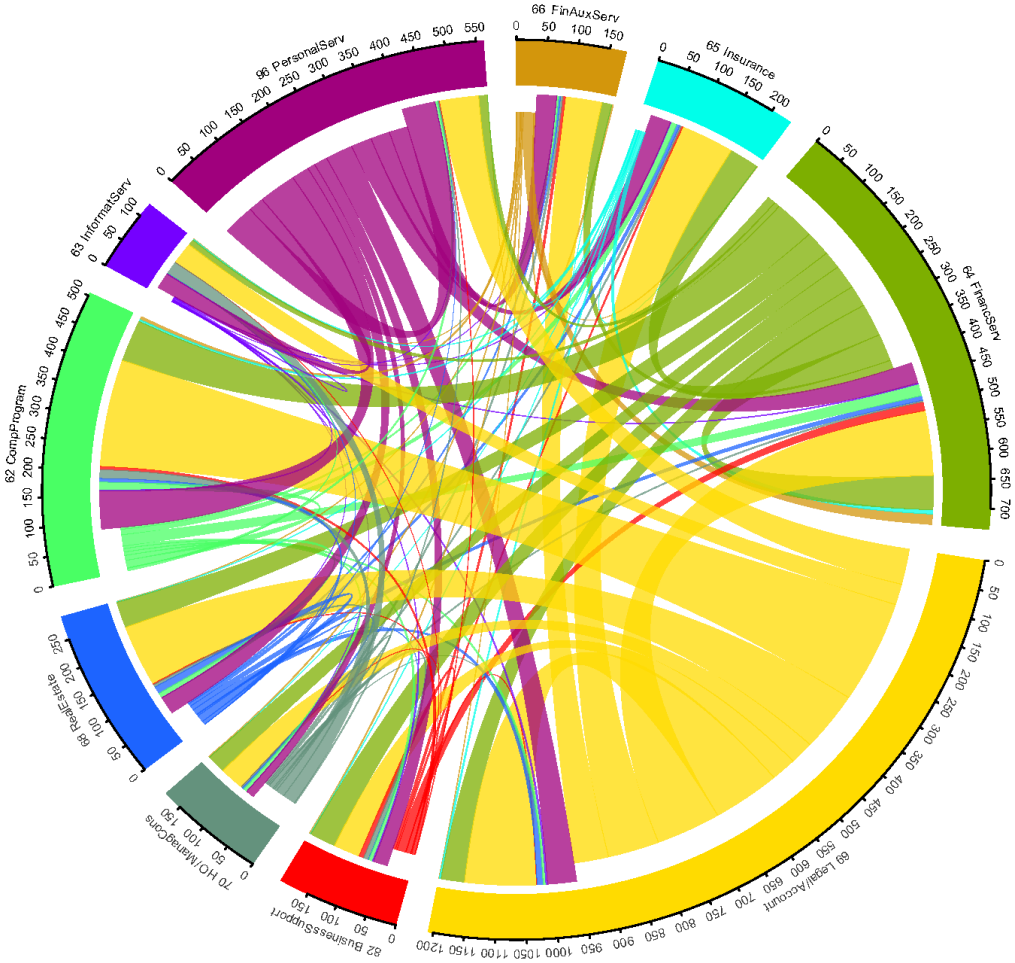
Notes: For each cluster, the table reports the number of links extracted from the webpages of companies that point to different type of targets.

While the network graph represents only a sample of targets, Table 11 reports the total number of targets by type and cluster. The table reports 42 links to higher education institutions extracted from the websites of financial companies based within the London cluster. However, the overall share of this type of target for this sector is much smaller than it is for the Digital-Health sector. In contrast, links to other financial companies are rather frequent across all clusters suggesting more frequent interactions between firms from this sector.

Despite most of the recovered financial companies belong to few different SIC classes; Figure 17 reveals interesting patterns of inter-sectoral connections. For example, “thick” flows of links generated by recovered companies from SIC 69 (i.e., Legal and Accounting Services) point to non-financial companies from SIC 62 (i.e., Computer Programming) and SIC 63 (i.e., Information Services). These connections may capture the emerging

integration between traditional financial services and the digital sector. Overall, Figure 17 confirms that inter-company connections are frequent within the financial sectors.

Figure 17: Web-links originated by financial companies across SIC industries



Notes: The circle plot shows the flows of web-links between different SIC industries. All links are generated by companies classified as part of the Financial Sector but they can redirect to companies outside this sector. The outer ring represents the ten most frequent industries from (and to) which we observe web-links. Labels on the external side of the outer ring refer to the number of links. The size of the flows crossing the circle increases with the number of web-links exchanged. Links generated by companies from (to) one SIC class are more distant from (closer to) the section of the outer ring for that class.

2.5 The Processing Industry

The objective of the analysis on the Processing Industry is to identify and map companies engaged in economic activities similar to the ones defining the North East Processing Industry Cluster. For this sector, we include in the restricted sample a small number of formal members of the NEPIC cluster organisation. The highly selective nature of the restricted sample generates a benchmark set of entities that are highly specific to the Processing Industry (see Table 12). Many of the entities extracted refer to chemical components (e.g., Pyridine, Polyketone, Biocide) or processes (i.e., X-ray crystallography, Thermal oxidation). The list of entities also includes a few concepts related to pollution control that are emphasized on companies' websites (e.g., Green Economy, Pollution

Prevention and Control). Based on this set of entities, recovered companies should be expected to belong mostly to the chemical and the pharmaceutical industries. However, Figure 18 shows that recovered companies with high W2V score (i.e., we retain only those at the top 5% with value label = 20) are distributed across a wide range of SIC codes. Wholesale firms constitute the greatest proportion of recovered companies, followed by firms included in a range of manufacturing SIC codes (i.e., Chemical Manufacturing, Other Manufacturing, Machine and Equipment Manufacturing). On the one hand, the wide dispersion of recovered companies across SIC codes suggests that the boundaries of the Processing Industry emerging from website data analysis are more inclusive than the one defined by SIC codes. On the other hand, this dispersion may signal a more serious misclassification problem for this sector.

Table 12: Benchmark sets of entities for the Processing Industry

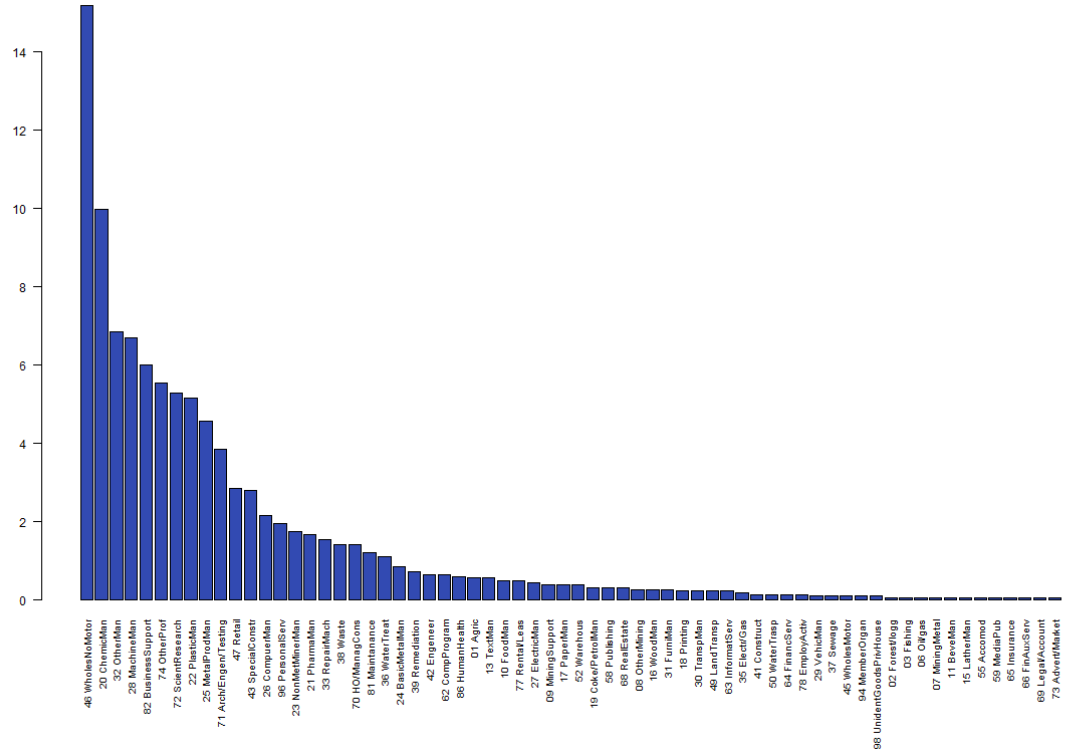
Chemical industry	Polichetoni	Polymorphism (materials science)	Refractory
Raw material	Methacrylate	Energy	Chemical process
Chemical substance	Polylactic acid	Capillary	Chemical synthesis
Organic chemistry	Thermal oxidation	Green economy	Thermosetting polymer
Pharmaceutical industry	Cyanation	Pollution Prevention and Control	Ester
Chemistry	Distillers grains	Engineering	Inorganic chemistry
Fine chemical	Pharmacy	Green nanotechnology	
Chemical engineering	Metallurgical assay	Waste management	
Pharmaceutical drug	Castor oil	Polyolefin	
Polymer	Special Obtain	Industrial wastewater treatment	
Sewage treatment	Polyol	Plastic	
Coating	Hydraulic fracturing proppants	Chemical reaction	
Speciality chemicals	Analyte	Solution	
Steel	Pyridine	Freeze-drying	
New product development	Analysis	Assay	
Biotechnology	Organic synthesis	X-ray crystallography	
Polimero	Acetic acid	Active ingredient	
Polyketone	PEEK	Contract research organisation	
Depyrogenation	Titanium dioxide	Biocide	
Docosahexaenoic acid	Waste	Gel	

Notes: The table lists all the entities from the extended benchmark set of the Processing Industry.

The geographical clusters obtained by running the DBSCAN algorithm with different parameters are listed in Table 13. For values of the radius parameter that are smaller than 15km we tend to obtain a very fragmented picture, with important urban areas split across multiple clusters. Instead, when we set the same parameter above 15km we obtain macro-

clusters due to ‘corridors’ of low-intensity sectoral presence linking higher density areas. Our favourite parameterization of the algorithm has *Eps* = 15km and *n* = 15. While these parameters are sufficiently restrictive to eliminate the noise generated by low density areas, they prevent excessive fractioning of urban areas in a number of different micro-clusters. In Table 13, the clusters identified by adopting this parameterization of the algorithm are highlighted.

Figure 18: Percentages of recovered companies by SIC code



Notes: The figure shows the percentages of recovered companies (y-axis) for the processing industry (only label = 20) by SIC code.

Figure 19 shows the map of the Processing Industry clusters identified by our preferred parameterization of the DBSCAN algorithm. Two agglomerations are identified in the Teesside Valley and they are named after the TTWAs of Newcastle and Middleborough and Stockton. These are the clusters from which we obtained the restricted sample. The main feature of the map is the dominance of two macro-agglomerations: one spreading around London (572 companies) and one with the centre in the Manchester TTWA (867 companies). DBSCAN identifies also smaller clusters in Scotland (Edinburgh and Glasgow) and in the South/South-East (Southampton, Cardiff and Bristol). We obtain Figure 20 by estimating the same parameterization of the DBSCAN algorithm while controlling for the overall density of non-Processing companies in the clustering areas. In the left-hand side panel we impose the central nodes to be located in areas with an overall density greater than the national average. Some of the smaller clusters previously identified in Figure 20 do not pass this test (i.e., Aberdeen, Edinburgh and Bristol). The modified algorithm still reveals a high concentration of companies related to the industry in the proximities of London and in the Midlands. On the contrary, none of the clusters passes the more stringent test in the right-hand side panel (i.e., density greater than twice

the national average). This result is mostly explained by the overall dispersion of recovered companies in many areas of the UK territory.

Table 13: Processing Industry clusters identified by the DBSCAN algorithm

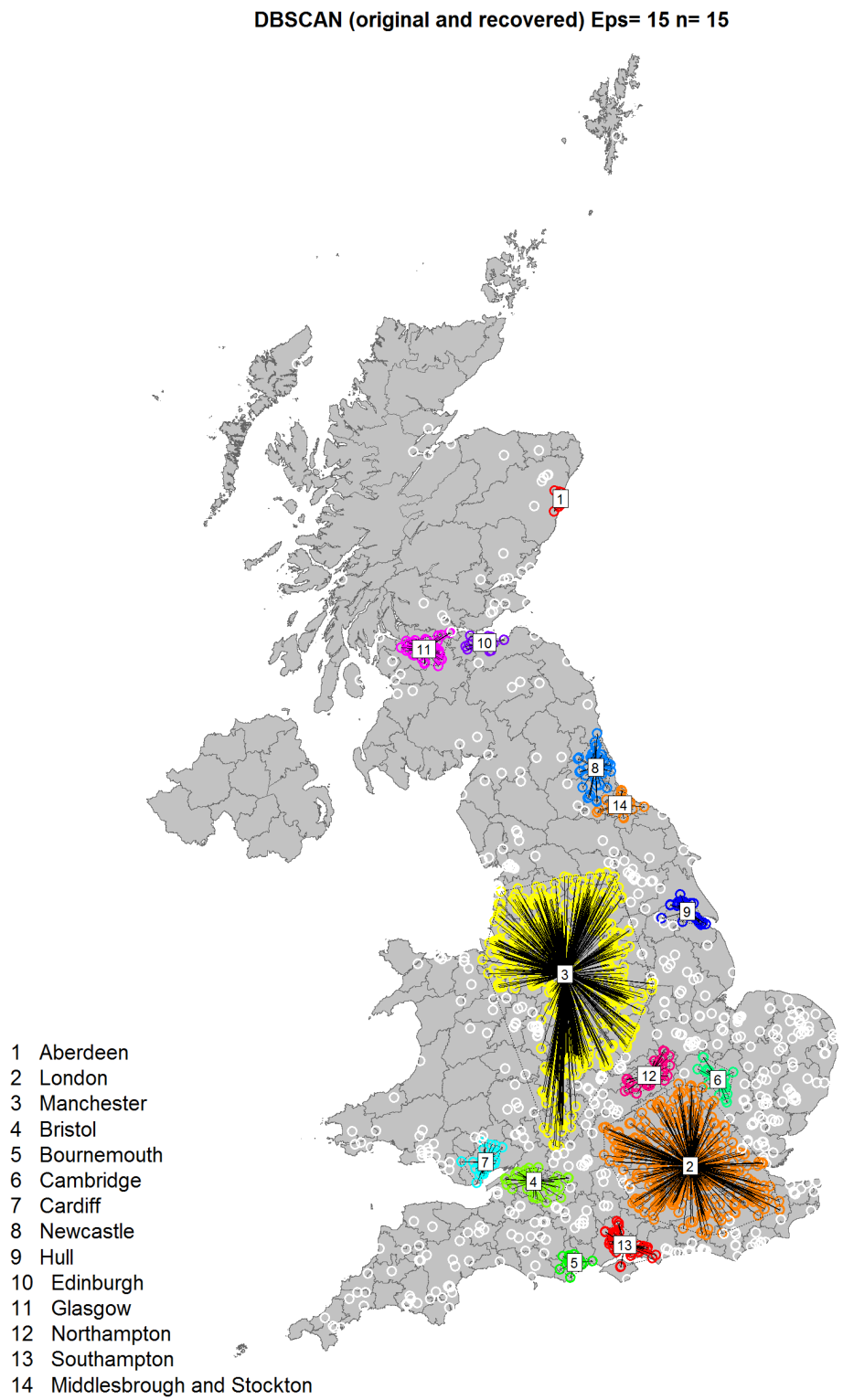
	eps=10km			eps=15km			eps= 20km		
	n=5	n=10	n=15	n=5	n=10	n=15	n=5	n=10	n=15
Aberdeen	22	22	22	22	22	22	27	27	27
Banbury	5								
Basingstoke		11							
Bedford		10							
Birmingham		138							
Blackpool	16	11							
Bournemouth	22	22	21	24	24	24		25	25
Brighton	7								
Bristol h	44	30	28		55	50			
Bury St Edmunds	5								
Cambridge	31	27	20			32			
Canterbury	8			10					
Cardiff	34	21	21	49	35	32			
Chelmsford	9								
Cheltenham		15	15						
Chichester and Bognor Regis	5								
Cinderford and Ross-on-Wye	6			6					
Coventry	16	10							
Crawley	22	22	21						
Dudley			115						
Edinburgh	19	15		20	20	20			23
Exeter				5			5		
Glasgow	35	32	32	53	51	50	79	75	52
Great Yarmouth				7					
Grimsby	13	12							
Guildford and Aldershot			10						
Hastings	7			7					

	eps=10km			eps=15km			eps= 20km		
High Wycombe and Aylesbury		9							
Huddersfield			76						
Hull	17	17	17	34	31	31		35	35
Ipswich	9				11				23
Kettering and Wellingborough		13							
King's Lynn	6			17	14			16	
Lancaster and Morecambe	8			8					
Leamington Spa	6								
Leicester	27	17							
Lincoln	5								
London	482	385	329	1818	1634	572	2062	1908	1777
Manchester	579	398	369			867			
Medway	42	36	25						
Middlesbrough and Stockton	31	23	22			29			
Milton Keynes	11	11							
Motherwell and Airdrie	16	10							
Newcastle	45	43	38	87	87	55	89	88	88
Northampton	42	12				42			
Norwich	15	10		22	17			25	23
Nottingham	219	199	47						
Oxford		27	17						
Pembroke and Tenby							5		
Peterborough	9								
Plymouth	7			8			18		
Preston			15						
Rhyl	5			6					
Salisbury	7								
Sheffield			37						
Shrewsbury	5								
Southampton	49	48	47		55	52			58
Southend		10							

	eps=10km			eps=15km			eps= 20km		
St Austell and Newquay							5		
Swansea	10	10			13			12	
Swindon	11								
Taunton							5		
Telford	11	11							
Trowbridge	5								
Worcester and Kidderminster			15						
Worthing	6								
Yeovil				8				6	
York	9								
Outside cluster	**343**	**676**	**1004**	**152**	**294**	**485**	**68**	**146**	**232**
Total	**2363**	**2363**	**2363**	**2363**	**2363**	**2363**	**2363**	**2363**	**2363**

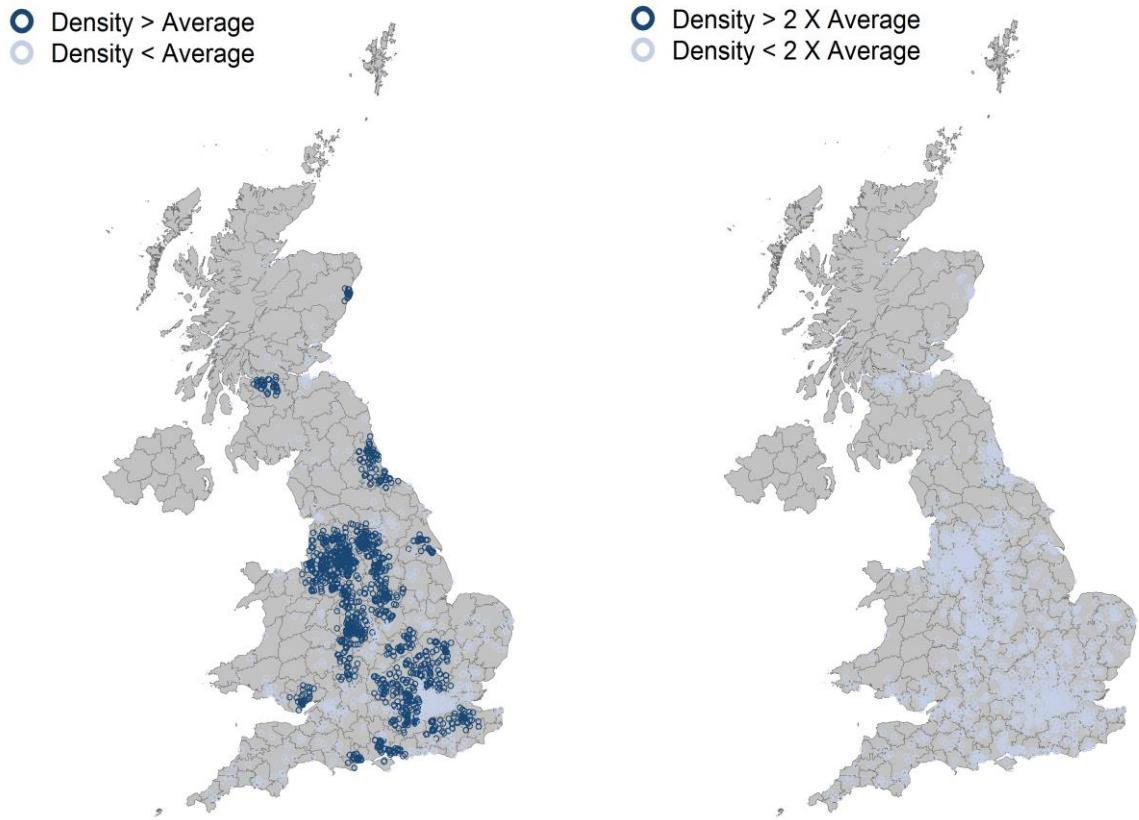
Notes: For each of the clusters identified by the DBSCAN algorithm, the table reports the number of original and recovered (only label=20) companies. Cells are left empty when the cluster is not identified for a particular combination of the *Eps* and *n* parameters. Clusters take the name of the Travel to Work Area where most of the companies belonging to that cluster are based. We highlight in the table the clusters that are more robust to different parameters of the DBSCAN algorithm.

Figure 19: Map of Processing Industry clusters



Notes: The figure shows the geographical location of the Processing Industry clusters (colors) identified by running the BDSCAN algorithm (*Eps* =15km, *n*=15). The map is obtained by locating geographically all original and recovered companies with label = 20. White circles on the map represent companies that are not assigned to clusters.

Figure 20: Processing Industry clusters after controlling for the overall concentration of non-Processing Industry companies in the clustering area



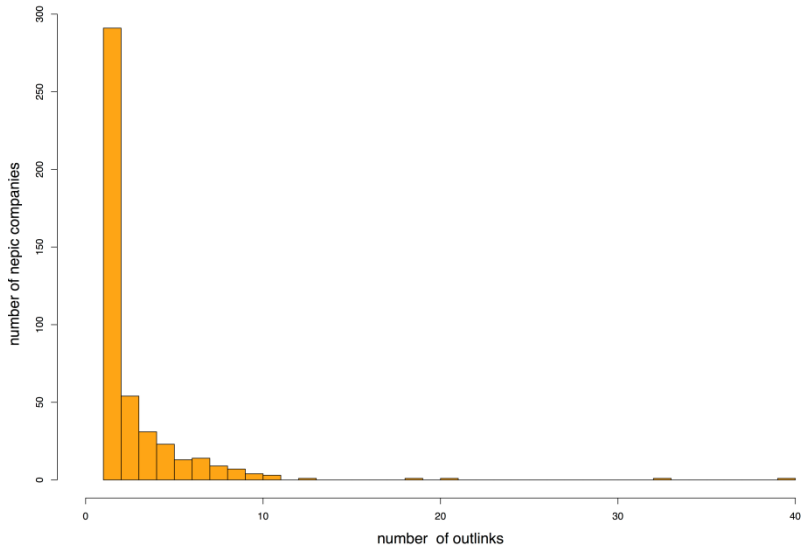
Notes: The two maps are obtained by running a version of the DBSCAN algorithm that accounts for the overall concentration of companies that are not part of the Processing Industry within the clustering area. Within the radius of the companies plotted in darker blue there is a relative density of companies from the Processing Industry (obtained as the number of Processing Industry companies within radius over the total number of companies within radius) greater than the national average (left-hand side panel), or greater than twice the national average (right-hand side panel).

Similar to the other two sectors, the distribution of companies by the number of links extracted approximates the shape of an exponential distribution with very high density at value one (Figure 21). The network graph in Figure 22 reveals that many companies include on their websites links to higher education institutions (yellow circles) or not-for-profit organisations (orange circles). Compared with the network graphs for the other sectors, the link network for the Processing Industry shows a number of small connected components (i.e., groups of vertices connected with each other's but disconnected from the rest of the graph). These components often include one or more companies from other sectors. This feature of the network may suggest closer functional relationships between small groups of companies. The more densely connected portion of the network includes also a number of ".gov" websites.

Overall, these features of the graph suggest that the Processing Industry is characterized both by frequent functional relationships between companies and other institutional actors and by collaborations across small groups of companies. When we consider this analysis at the level of the individual clusters (Table 14), we find that companies belonging to the

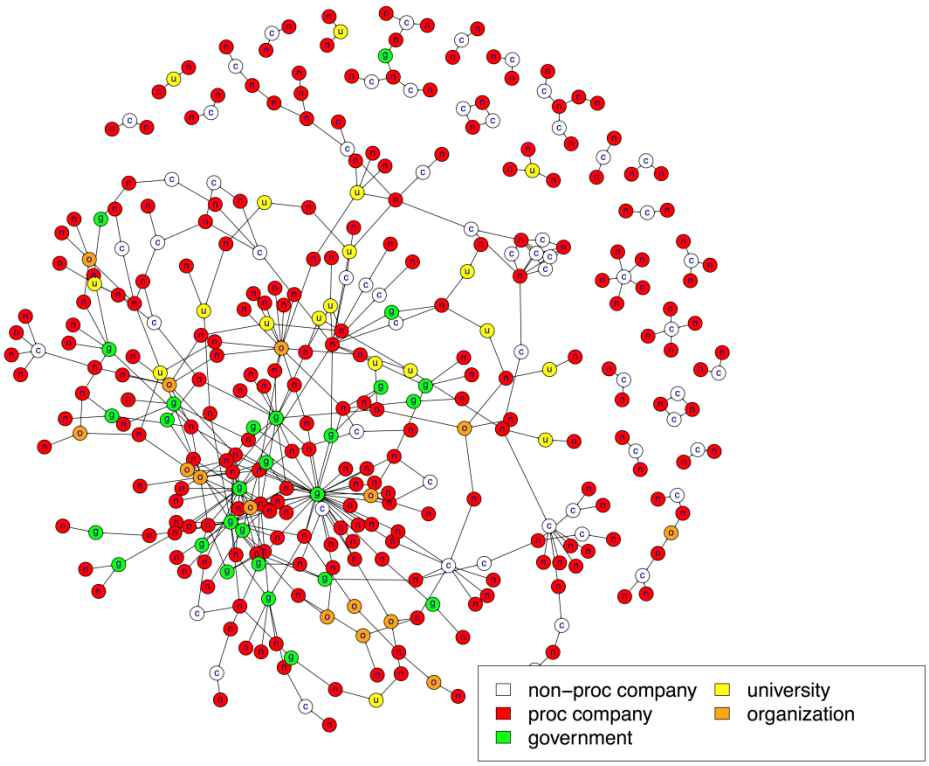
geographical clusters of London and Manchester generate most links to government websites, and to the websites of not-for-profit entities.

Figure 21: Distribution of Processing Industry companies by number of outlinks



Notes: The figure shows the distribution of Processing Industry companies with outlinks by the number of outlinks extracted.

Figure 22: Network graph of the web-links originated by Processing Industry companies



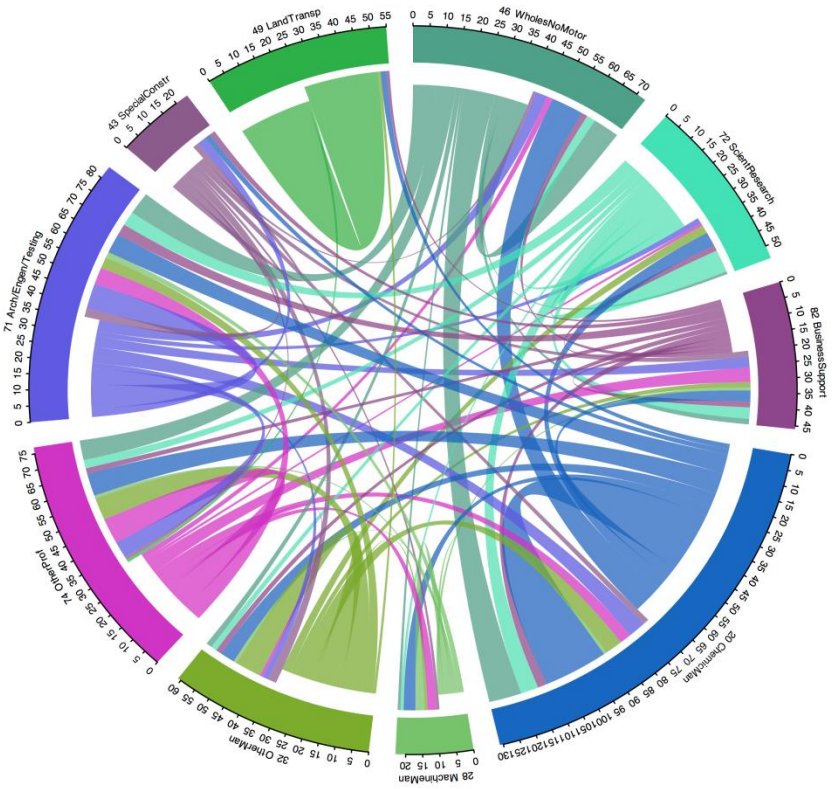
Notes: The network graph represents the links that from the websites of Processing Industry companies redirect to the websites of other companies (either from the same sector or not) or institutions.

Table 14: Number of links by type of target (all sources are Processing)

	company	government	organisation	university
Aberdeen	8	0	1	0
London	231	58	35	39
Manchester	252	66	32	22
Bristol	8	0	3	1
Bournemouth	13	0	0	0
Cambridge	23	3	6	8
Cardiff	20	0	1	1
Newcastle	14	5	1	2
Hull	5	2	2	2
Edinburgh	4	2	1	3
Glasgow	20	10	1	4
Northampton	8	8	8	4
Southampton	24	6	2	4
Middlesbrough and Stockton	4	1	1	0
Not in cluster	210	46	20	46

Notes: For each cluster, the table reports the number of links extracted from the webpages of companies that point to different type of targets.

Figure 23: Web-links generated by Processing Industry companies across SIC sectors



Notes: The circle plot shows the flows of web-links between different SIC industries. All links are generated by companies classified as part of the Processing Industry but they can redirect to companies outside this sector. The outer ring represents the ten most frequent industries from (and to) which we observe web-links. Labels on the external side of the outer ring refer to the number of links. The size of the flows crossing the circle increases with the number of web-links exchanged. Links generated by companies from (to) one SIC class are more distant from (closer to) the section of the outer ring for that class.

For this sector, the pattern of web-links reveals a greater dispersion of targets across SIC classes than for the previous sectors. Indeed, the flows represented in Figure 23 are generally thinner than the ones in the previous circle plots. This evidence suggests that companies in the Processing Industry are more likely to establish functional relationships with companies from a wider range of sectors. In particular, engineering/testing companies (SIC 71) have a portfolio of incoming and outgoing links that is very diversified by SIC code of the origin or target companies. However, there are two noticeable exceptions to this pattern. First, companies classified as Land Transport (SIC 49) are almost exclusively linked to companies from the same SIC class (20 out of 25 outgoing links point to companies within the same SIC class). Second, about one third of the flows from Chemical Manufacturing (SIC 20) points to other companies in the same SIC class. The different likelihood to which firms classified under different SIC codes include on their websites links to companies from their own or from different SIC codes is an interesting finding. However, caution is required when interpreting this result. First, the sample of companies for which we obtain out-links is a selected subset from the population. Second, we have not yet developed a reliable classification of the functional relationship expressed by individual links. Qualitative analysis of individual companies' websites or the development of an algorithm that identifies the nature of individual links (i.e., clients, suppliers, partnerships) is required before interpreting these findings.

3. Qualitative analysis

While the quantitative analysis investigates each sector at the national level, the qualitative component of the study is structured around one case study of geographical cluster for each sector.

The primary aim of the qualitative study is to examine the nature of relationships - both within and across sectors - inside the cluster, the nature of knowledge transfer and spillovers between firms in the cluster, and the benefits and obstacles to firm growth of locating in clusters. In order to gauge the strategic importance of relationships within the cluster in the context of the increasing internationalisation of economic activity the case studies aim to collect qualitative data about the strategic importance of linkages between the key firms and stakeholders in each cluster, their partners outside the cluster and their international partnerships and networks. Qualitative analysis is also used to test the potential and the limitations of the quantitative analysis based on website data.

A common questionnaire was developed for the three clusters to enable consistency in the data collection but it was also flexible enough to enable industry and regional-specific differences to be discussed in detail. Interview topics included questions about the nature of the organisation's relationships to other organisations within the cluster, outside the cluster and region, as well as links to institutions. Interviews were recorded and coded to identify research themes across the case studies. What follows is a list of the main topics explored during the interviews:

- The identification of the key strategic partnerships within the cluster (i.e., other firms in the same industry, firms from other industries, universities or technical colleges, specialised business services).
- The strength and importance of industry and cross-sector linkages between various stakeholders in the cluster.
- The specific function of firms' value chain supported by the cluster.
- The role of cluster organisations such as universities, local training colleges, local knowledge intensive business firms (KIBS) in the transfer of knowledge and best practice within the cluster.
- The strategic importance of relationships within the cluster compared to relationships outside the cluster and with international partners.
- The sources of industrial dynamisms within the clusters and barriers to growth.
- The role of government (local, national) in supporting cluster development.

Throughout the exposition of the qualitative results we frequently use the following concepts:

- **Concentration of firms:** refers to the concentration of related firms in a geographical area but it does not imply interactions between them.
- **Cluster of firms:** Concentration of related firms that cooperate as well as compete. This concept is closely related to Porter's definition of a cluster (see Introduction).
- **Cluster organisation:** Local organisations whose role is to foster links and networking between firms in an area.

The discussion of the three case studies is organised around the common themes explored in the questionnaire. Given the differences in the nature of the industries and geographical areas studied, the specific methodologies of each case study are discussed in each sub-section. Of the three case studies, the Digital Health industry in Birmingham was the only one conducted after the completion of the classification exercise based on website data analysis. Therefore, this case study is both an opportunity to test and complement the quantitative results. More specifically, to test the quantitative classification results, we collected qualitative data for a sample of firms that were classified as being part of the Digital-Health sector. In addition we investigate through interviews the nature of their links to other local firms and organisations.

3.1 North East of England Process Industry Cluster (NEPIC)

3.1.1 NEPIC: Historical Development of the Cluster

The chemical process industry in Teesside was dominated by ICI and the region developed a strong reputation for chemical manufacture. The restructuring of the company during the 1970s led to an increased proportion of employment being offshored (Clarke, 1985) and diversification towards higher value products in specialist chemicals and pharmaceuticals (Chapman, 2005).

The decline of ICI's presence on Teesside and purchase of assets by other firms resulted in the fragmentation of the industrial complexes established by the company (Chapman, 2005). There remain significant legacy assets in the area but the region has suffered a negative reputation associated with the demise of ICI (Chapman, 2005). The Teesside Chemical Initiative was established in 1995 to utilize the existing asset base, both the physical infrastructure linkages between plants and knowledge capabilities, to support diversification into other industries and for "...stabilizing rather than developing the complex" (Chapman, 2005:606).

The North East of England Process Industry Cluster (NEPIC) was established in 2004. Two existing trade bodies were merged: Teesside Chemical Initiative, which represented heavy chemical industries; and Pharmaceutical and Speciality Cluster, which represented pharmaceuticals and life sciences (NEPIC, 2105a). The aim of establishing NEPIC was to connect these industries and to develop a presence in the global market, which had been lacking since the decline of ICI on the site (Higgins, 2013). NEPIC is the cluster management organization. The aim of the organization is to increase economic development and international presence, including attracting inward investment and

developing supply chain relationships (Higgins, 2013; Lammer-Gamp et al., 2014), to support the sustainability of chemical process industries in the region.

The chemical processing sector in the region represents £10billion of gross domestic product (GDP) and was expected to invest £7billion into the region by 2015 (House of Commons, 2009). NEP1C has been successful in establishing European recognition and was the first UK cluster organization to become an EU Accredited Cluster Management Organization (Higgins, 2013). NEPIC has also transitioned from an initially publically funded organization to become privately funded in 2011 (Higgins, 2013). Key income streams for NEPIC are membership fees, grant-based projects (UK- and European-based) and fee-based consultancy services. NEPIC, the cluster management organization, has generated £3.0billion of gross value added (GVA) for the north east regional economy since 2003, secured 83 significant investments and 4,500 jobs (secured or created) (NEPIC, 2015c).

3.1.2 NEPIC: Case Study Sampling Methodology

The focus of the case study is the formal cluster management organization, referred to as NEPIC throughout the remainder of the report. NEPIC has identified 720 participating organizations in the cluster, which includes 340 paid members (based on the 2015 directory). Participating organizations differ to members because they do not pay an annual fee. Membership provides access to NEPIC personnel and involvement in activities. Participating organizations are those that have been involved with NEPIC activities (such as specific programs run by NEPIC) and are able to access publically available resources (web-based). The number of participating organizations is the key benchmark recognized by the European cluster accreditation as this reflects the organization's scale and scope.

A desk-based scoping exercise has been undertaken to construct an understanding of the history, function and scale of the concentration of chemical processing firms in the region. In addition, a site visit to the cluster management organization (NEPIC) has been undertaken and an interview with a local policy-making organization has further informed the identification of key stakeholders and research findings.

Three target groups have been identified through the scoping exercise (Table 15):

1. Registered fee-paying members of NEPIC, which includes private and public organizations in the manufacturing and service sectors, as well as institutions;
2. Participating organizations (who may or may not be registered members but use the services of the cluster organization) and include the range of organization types identified in the member sample above;
3. Innovation support institutions, which include national centers of excellence, research centers and universities.

These groups reflect the diversity of organizations and methods of engagement undertaken through NEPIC. Participants were selected using a random sampling approach to avoid selection bias. The NEPIC directory provides a full listing of members and key contact details (publically available). To access the participating organization group, a sub-group of organizations that have participated in a significant and recent programme were targeted. The Business Acceleration for SMEs (BASME) program was funded by the

Regional Growth Fund to support growth for SMEs in the region (Ford, 2015). This databank provides access to recently participating firms. The final target group are innovation support institutions, which have been identified by NEPIC as key assets for the cluster. In addition to the random sampling, a snowball approach was used to identify key informants in each of the target groups and support access to participants. Together these approaches have enabled a comprehensive sample of NEPIC participants beyond defined SIC listings. A breakdown of population samples used to construct the sample database is provided in Table 15 below.

Table 15: Population sampling

Target group	Population	Number of invitations	Number of interviews (response rate)
Scoping study	3	3	2
Registered members *Source: NEPIC (2015a)*	340	56	9
Participating organizations -BASME programme participants *Source: NEPIC (2015a)**	416	11	1
Innovation support *Source: NEPIC (2015b)*	22	8	3
Total		78	15 (19%)

Notes: *Restricted sample due to limited contact information for individual organizations and overlap with membership listing.

In total 15 interviews have been conducted and the overall response rate from invitations to participate in the study was 19%. The sample of interviewees includes both members and non-members of NEPIC (although all have had engagement with NEPIC). Organizations include manufacturing- (4), service-(3) and energy-related (3) industries, which are both small and medium sized enterprises (SME) (7) and multinational enterprises (2). Innovation support institutions (3) were also SMEs. Participating organizations and interviewees have been anonymized and an interviewee code used to describe the interview (Table 16).

Table 16: Summary of the interviewee sample

No.	Interviewee Code	Interviewee Position
1	Scoping 1 (private)	CEO
2	Scoping 2 (public)	Head of Strategy
3	Non-member-SME-Energy	Director
4	Member-SME-Energy	Director
5	Member-Utilities	ex-Business Development manager
6	Member-SME-Manufacturer 1	Business Development Manager
7	Member-SME-Manufacturer 2	Managing Director
8	Member-Multinational-Manufacturer 1	Site Manager & HR*
9	Member-Multinational-Manufacturer 2	Operations manager
10	Member-SME-Business Services 1	Managing Director
11	Member-SME-Business Services 2	Director
12	Member -SME-Business Services 3	Director
13	Innovation-Institution 1 (member)	Director
14	Innovation-Institution 2 (non-member)	Program Manager & Technical Bid Manager*
15	Innovation-Institution 3 (member)	CEO

Notes: * indicates that interview was conducted with multiple interviewees.

The websites of the participating organizations include affiliations with other network and sector-based organizations. The affiliations demonstrate active participation in the 'business community' and are a marketing tool. One interviewee (Member-Multinational-Manufacturer 1) stated that he was unaware of these affiliations and suggested that inclusion on the website was a 'marketing exercise'. Commercial partners are included in four of the ten company-based organizations. These business links are demonstrated mainly through case studies of work undertaken with prominent clients or the inclusion of logos of key actors in the sector to illustrate status in the market place. Key suppliers are listed only for the participant that is a distributor of products, again representing status in the market place through supplier brands. There is very limited information for site activities or functional relationships in large group-based websites (e.g. multinational organizations), franchises and micro-enterprises. When interviewees were asked about the representativeness of information provided on the website all stated that there was a time-lag in information presented.

3.1.2 NEPIC findings

The case study findings are summarized in Table 17 below. The overall findings illustrate the diversity of organization types and levels of engagement with NEPIC. The results also illustrate the variety of registered SIC codes for the interviewee's organizations, which includes manufacturing, distribution, consulting and other professional, scientific or technical activities.

Structure of NEPIC

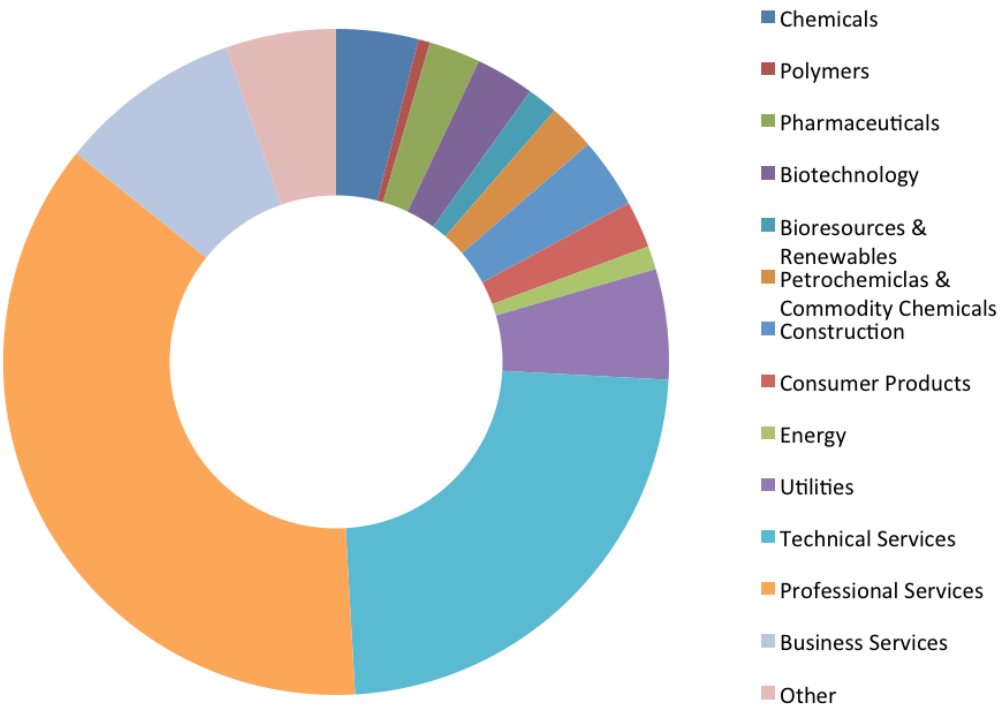
The concentration of firms in the region is a diverse set of inter-connected industries that form part of or engage with the chemical processing industries. NEPIC (the cluster management organization) have identified key industrial groups: chemicals; petrochemicals; speciality chemicals; polymers; pharmaceuticals; bioresources & renewables (NEPIC, 2015c). Organizations located in the concentration may or may not be participating organizations of NEPIC. The study has drawn on NEPIC datasets to identify participating organizations and therefore the representativeness of participating organizations in NEPIC to the wider population of chemical processing-related organizations cannot be ascertained in this study.

The diversification of industrial groups reflects the nature of the chemical processing sector and generates a significant challenge in mapping the industries through the standard industrial classification (SIC) system. Manufacturers may be registered as consumers or self-register in their customer market sectors rather than as a chemical processing industry. In addition, the 'Head Office Effect' (where local or regional business operations are reported through their head office or sales offices outside the region) is significant and can displace economic output to other regions of the country, notably London, where activity is reported (Scoping 1). These effects are apparent in the cluster mapping exercise, which has identified London as having a very significant concentration and the north east a relatively small concentration of organizations (see Figure 24). As a result, the significance of the sector for the north east regional economy is distorted.

Member organizations of NEPIC also undertake a wide range of activities (see Figure 24 below). These include manufacturing, bio-energy, logistics, construction and knowledge intensive business services. There is a high proportion of service activities (68.8% of activities are professional, technical or business services) and only 13.7% of activities relate to industries defined as those the sector represents (chemicals, polymers, pharmaceuticals, biotechnology, bioresources and renewables, petrochemicals and commodity chemicals) (NEPIC directory, 2015: based on reported activities that have be grouped by the author).

In order to identify more precisely the Processing Industry, during the quantitative analysis we selected only some of the SIC codes represented in the NEPIC cluster within the restricted sample. However, those codes do not fully represent the activities of NEPIC members which include many service-based companies. In addition, the selected SIC codes do not reflect the evolution of the activities and industries within the concentration of firms. Biotechnology, bioresources and renewables are emerging industries within the concentration of firms.

Figure 24: Breakdown of key activities of NEPIC members



Notes: Source: NEPIC (2015a). The categories reflect specialization defined by NEPIC, which have been aggregated by the author for analysis. Categories are not mutually exclusive and organizations may be listed under multiple activity groups.

Origin and description of the NEPIC cluster

All interviewees differentiated between the concentration of firms and the cluster organization (NEPIC). The concentration of firms includes two groups: (1) the core group of chemical processing firms and legacy assets from ICI that are primarily based at Teesside and represent a concentration of physical assets; and (2) the network of firms and organizations in associated industries in the wider north east region, which represents a concentration of firms in related industries. A summary of interviewee descriptions of these groups is provided in Table 18 below.

The core group is the concentration of assets and plants in the geographic area, described by one interviewee as a group of 6-7 key firms (Member-Multinational-Manufacturer 2). ICI played a significant role in developing the integrated infrastructures and assets in this concentration at Teesside. This core group of plants and assets had a high level of physical integration historically as inputs and outputs moved between plants (for example, through pipeline connections as well as infrastructure links). Over time, with the breakup of ICI and the closure of key plants, the level of integration has reduced. New firms that have moved to Teesside have not necessarily integrated with the physical infrastructure between plants (Member-Multinational-Manufacturer 2).

Table 17: Summary of interview results

Interviewee	Registered SIC code[1]	How the organization describes itself	Date of incorporation[1]	Employees[2]	Describes itself as 'part of the cluster'	Engagement with NEPIC
Non-member-SME-Energy	74901 - Environmental consulting activities	Business support - low carbon in the built environment commercial projects	2012	unknown	No	BA SME programme participant
Member-SME-Energy	05101 - Mining of hard coal from deep coal mines (underground mining)	Project development - energy	2010	unknown	Yes	Long standing member of NEPIC
Member-Utilities	unknown	Consultant - construction and engineering design	unknown	unknown	Yes	Long standing member of NEPIC
Member-SME-Manufacturer 1	28131 - Manufacture of pumps	Manufacturer and distributor of industrial equipment	1954	25	Yes	BA SME programme participant Long standing member of NEPIC
Member-SME-Manufacturer 2	46750 - Wholesale of chemical products	Distributor - chemicals	1981	100+	Yes	Long standing member of NEPIC
Member-Multinational-Manufacturer 1	20150 - Manufacture of fertilisers and nitrogen compounds	Manufacturer - chemicals	1997	600	Yes	Member but limited relevance
Member-Multinational-Manufacturer 2	20590 - Manufacture of other chemical products n.e.c.	Manufacturer - chemicals	1999	80	Yes	Long standing member of NEPIC
Member-SME-Business Services 1	62090 - Other information technology and computer service activities	Creative design agency (design, branding and website	2008	unknown	Yes	BA SME programme participant Subsequently became a member of NEPIC

Interviewee	Registered SIC code[1]	How the organization describes itself	Date of incorporation[1]	Employees[2]	Describes itself as 'part of the cluster'	Engagement with NEPIC
		development)				
Member-SME-Business Services 2	74909 - Other professional, scientific and technical activities (not including environmental consultancy or quantity	Consultant - process engineering	1988	8	Yes	BA SME programme participant Long standing member of NEPIC
Member -SME-Business Services 3	70229 - Management consultancy activities (other than financial management)	Operational HR for small businesses	2015	1	No	Member of NEPIC for single year
Innovation-Institution 1 (member)	72190 - Other research and experimental development on natural sciences and engineering	Strategic consultancy for government, business and academic communities	2003	12	No	Reciprocal membership with NEPIC
Innovation-Institution 2 (non-member)	74909 - Other professional, scientific and technical activities (not including environmental consultancy or quantity surveying) n.e.c.	Business support and product testing	2003	unknown	No	Attended occasional event only
Innovation-Institution 3 (member)	74909 - Other professional, scientific and technical activities (not including environmental consultancy or quantity surveying) n.e.c.	Technology, innovation and business development support	1989	45	Yes	Long standing member of NEPIC

Notes: Source: [1]FAME database (2015); [2]NEPIC (2015a).

Table 18: Summary of interviewee description of the cluster

Interviewee	What is the cluster?
Non-member-SME-Energy	"My interpretation of what NEPIC membership means is either university or a company, commercial company that makes widgets and stuff for the process sector."
Member-SME-Energy	"...we kept getting confused whether NEPIC represented the cluster or NEPIC was the cluster. You know, because its name is North East Process Industry Cluster. I tend to think of NEPIC as the organization that represents the cluster and does things for the cluster and provides all sorts of very valuable services and all of that. Other people say no NEPIC is what I used to think of as the ICI chemistry set-up in the north east, you know, it's the cluster of plants."
Member-SME-Manufacturer 2	"Well, it's just a network of, network of industries within the supply chain for the chemicals and pharmaceuticals. That's really what the cluster is...The companies located in that area specializing in those fields and the interests."
Member-Multinational-Manufacturer 1	"There's two sides to NEPIC if I was honest. There's obviously the legacy chemical companies - large, single stream plants, manufacturers on Teesside. But then obviously NEPIC does represent a broader spectrum of process industries in the north east. So you know, in terms of pharmaceuticals, batch processes, fine organics, those sort of people. So there's probably two, in reality there's probably two sides to it."
Member-Multinational-Manufacturer 2	"If we weren't members of NEPIC we would still be part of the cluster. We just wouldn't be members of the cluster organization...The cluster is the group of companies. Because that's what drives the, it's no good having a cluster organization if you've only got two companies." "There's a group of companies in NEPIC...If you like, they're the bigger of the ex-ICI, mainly ex-ICI companies, the bigger ones, that form if you like, the core of the group, or the critical mass. And that's maybe six or seven of those."
Member-SME-Business Services 2	"Well, I think the staff in the office [NEPIC] bring together a number of skills and a number of different experiences and backgrounds. But also when you get together with other members that is extremely useful because you know, the networking opportunities, just seeing members and, well, meeting new people but also resurrecting acquaintances from the people you've met before. That's always extremely useful."
Innovation-Institution 3 (member)	"It is in my view one of the best examples of an industrial cluster in the country, never mind the region. And it is held together partly by the personality of [the CEO], who is a larger than life character and has managed to build around him a lot of loyalty."

The second group of firms is the network of organizations in associated industries related to this core (pharmaceuticals, batch process, fine organics, and equipment). Related service industries, including technical, professional and business services are also part of this group. These organizations represent a concentration of firms in related industries but

there is no implied interaction or co-location benefit between them other than networking. The diversity of this secondary group makes it difficult to define the group more specifically based on the interview data.

NEPIC - The cluster organization

The cluster management organization (NEPIC) provides a critical function to the activities of the concentration of firms. NEPIC undertakes activities on behalf of the organizations (Member-SME-Energy). It is distinct from the organizations themselves and provides an overarching role to facilitate interaction between members. One interviewee described the role as partly about substituting for a large anchor organization (the role partially undertaken by ICI before its breakup) (Member-SME-Energy). The fragmentation of the industry at Teesside has meant that no single organization undertakes long term, strategic planning or investment activities that are important for organic growth across the area. NEPIC attempts to bring together organizations to collectively achieve this through targeted projects and initiatives and acts as coordinator for these types of activities. An example of this is the current ‘re-integration’ project underway that is developing an understanding of where potential opportunities for physical integration exist for organizations in Teesside.

Consciousness of being part of a cluster

Consciousness of being part of a 'cluster' differed between organizations approached for the study (Table 17). The interviewee descriptions illustrate the overlapping definition of 'cluster' to organizations (Table 18). Most interviewee's identified as being part of the 'cluster' (9) however, three organizations that had directly engaged with NEPIC (two members and one participating organization) did not consider themselves to be part of the 'cluster'. These organizations did not consider themselves to be 'embedded with other organizations in the cluster' (Innovation-Institution 1 (member)) or a 'traditional fit' (Non-member-SME-Energy; Member -SME-Business Services 3) citing that their operations were not directly related to the manufacturing specialism of the 'cluster'. One interviewee (Member-Multinational-Manufacturer 1) identified as being part of the Teesside (core) concentration but had limited relevance to the secondary group of organizations or NEPIC. Requests for interviews were declined by three members of NEPIC because they felt they were not part of the 'cluster' having not had any interaction with other members. This preliminary indication has identified a group of paid members that are not actively participating or conscious of their involvement. The scale of this group cannot be ascertained from this study.

Cluster and region reputation

The core concentration of firms is spatially tied. Although participating organizations are located in the wider region, and in some cases nationally (member organizations may have a relationship with NEPIC for 'information gathering exercises' through the head office rather than plants located in the region), the core group are located around Teesside and are geographically concentrated. Teesside and the North East region have a reputation for chemical processing externally. This reputation has been built from the legacy of ICI as a key player in the industry. Since the breakup of the company it is more difficult for organizations seeking to work with firms in the sector to identify and access them (Member-SME-Energy). NEPIC has built a reputation as a cluster management

organization, which is reflected in the European accreditation for cluster management and Dr Higgins accreditation for cluster leadership. These accreditations help signpost the cluster management organization as the umbrella organization that provides a route into the individual firms in the concentration.

Linkages with other firms in cluster

There are two forms of linkage cited by interviewees: (1) linkages based on geographical proximity; and (2) linkages based on engagement between NEPIC members. Linkages based on the concentration of firms were identified for a subset of interviewees (multinational manufacturers with world-scale plants for commodity chemicals). Two of the interviewee's cited specific relationships with members in their supply chains based on integrated infrastructure in the area (Member-Multinational-Manufacturer 1; Member-Multinational-Manufacturer 2). Access to infrastructures does not determine the location of these plants but is a benefit compared to alternative locations. The scale of integration between manufacturing companies has reduced as firms have closed and one interviewee noted that new firms do not necessarily integrate despite potential benefits (Member-Multinational-Manufacturer 2). This form of integration demonstrates a cluster aspect within the concentration of firms, although for a very small and geographically specific group.

Functional linkages between NEPIC members were difficult to identify. Interviewees stated that they do have trading relationships with other firms located in the region, which may or may not be members of NEPIC. These relationships were not directly attributable to proximity. One example was found of strategically important trading relationships that were directly established through participation in NEPIC events. Member-SME-Manufacturer 1 has been able to build relationships with two large customers that were geographically proximate (within 10 miles) through attendance at a networking event. These customers had previously been difficult to access and an introduction was facilitated through NEPIC.

Relationship with NEPIC cluster management organization

Interviewees stated that their relationship with the cluster organization was also important (Innovation-Institution 1 (member); Member-SME-Business Services 2; Member-SME-Manufacturer 2). NEPIC facilitates collaboration between members by acting as the 'trusted partner' (Member-SME-Energy; Member-SME-Business Services 1). Confidential benchmarking is undertaken by NEPIC to identify collaborative opportunities, best practice activities and improve the overall competitiveness of firms within the cluster. The expertise of NEPIC employees, all of whom have a background in senior positions within the chemical processing industries, enables informed, tailored and relevant support to enrich the capacity of firms within the concentration to expand. It is a benefit to have an independent administrative group so companies feel confident sharing information that may be commercially sensitive (Member-SME-Manufacturer 2). As a result, linkages with the cluster management organization were noted as significant in developing linkages between member organizations.

Linkages with firms outside the cluster

It was difficult to determine specific external linkages between firms as interviewees stated that these were vast and specific details could not be provided. Several interviewees did identify NEPIC as having a role in strengthening international presence. The cluster management organization has a strong presence in the region and internationally. NEPIC actively build relationships with clusters and trade associations in overseas regions (in particular, India) to support internationalization of NEPIC members, mainly SMEs. It also has a strong presence as a leading cluster management organization in Europe, which helps attract inward investment and participation in European initiatives. This can be a benefit for both NEPIC member organizations and related-chemical processing organizations in the region because the industries and firms are relatively 'invisible'. The complexity of the chemical process industry and its role as an industry at the foundation of many value chains results in limited awareness of the scale and importance of the sector for the regional and national economy. NEPIC provides a voice, presence and marketing resource for firms and the sector more widely. Interviewees reported that links between NEPIC and organizations are largely site-only and sister sites would use their equivalent regional body.

Linkages with institutions (non-other firm in cluster): Importance of links and types of linkages

NEPIC has identified 22 innovation institutions that it currently has links with to support organizations in the concentration of chemical processing firms (NEPIC, 2015b). These include 12 national centres of excellence, 5 research centres and 6 universities based in the region. Individual organizations interviewed stated that they access these institutions, and others, directly. Links to local universities for graduate skill intake (Member-Multinational-Manufacturer 1), research projects (Non-member-SME-Energy; Innovation-Institution 3 (member)) and technical testing facilities (Member-SME-Manufacturer 2) were cited but the relationships were not viewed as strategically significant. The North East Chamber of Commerce (NECC) was repeatedly cited as another route for organizations to engage with. Interviewees stated that the NECC provided more international links and global opportunities than NEPIC, which has greater value for regional issues (Member-SME-Manufacturer 1).

Added value for firms that are part of NEPIC?

Industrial Leadership

NEPIC is industry-led and industry-focused. Industrial leadership was cited as a key factor in its success because it ensures deeper sector knowledge of firms in the concentration that teases out new innovations, products, processes and investment (Member-Utilities). This 'industrial intimacy' was identified by the same interviewee as a factor in the economic growth of the region, stating:

> *"[t]his growth in industrial activity is unlikely to have occurred without the industrial leadership in this region - and its industry thrust teams North East Bioresources & Renewables and Process Industry Carbon Capture and Storage Initiative - promoting nationally and internationally the infrastructure and engineering capability of Teesside"* (Member-Utilities).

Industry members are integrated into NEPIC by setting and delivering initiatives. Industry engagement is structured through membership of the NEPIC leadership team and also involvement in theme-based working groups (thrust teams). By devolving the management and undertaking of some activities to industry participants it increases the sustainability of the cluster organization (Innovation-Institution 3 (member)). Some projects require funding, either from companies directly or NEPIC-sourced funds (through UK government or other funding mechanisms). Sourcing funding for projects is a key activity for NEPIC.

Industry Voice

NEPIC provides a single voice for the chemical processing industry. Although other networks and organizations also act as an industry voice, NEPIC has the scale and critical mass of organizations to lobby policy-makers and build an international presence to attract inward investment (Member-SME-Manufacturer 2). Two interviewees, both representatives of multinational firms with sites in the core concentration, identified limitations to NEPIC's role as a voice for the industry. One interviewee reported that NEPIC's role as an industry voice is in conjunction with the local enterprise partnership who also lobbies on behalf of industry (Member-Multinational-Manufacturer 1). The second stated that NEPIC mainly adds value for regional issues (Member-Multinational-Manufacturer 2).

Business Opportunities – Networking

The development of business opportunities through networking had a mixed response from interviewees. Interviewees identified networking events as providing the opportunity to meet a large number of organizations. Examples of different forms of networking are outlined in the first four rows of Table 19. However, another group of interviewees had not developed strategically significant business opportunities through the cluster (last two rows of Table 19).

The cases illustrate the different expectations to NEPIC membership by individual firms. One of the firms stated that they have not used the NEPIC network to identify business opportunities and another that the network did not generate significant referrals. The majority of interviewees (7/10 participating organizations) were members of multiple industry-based networks or clusters to increase their presence in the market and community, although the defined benefit from each varied (information gathering from industry-specific networks, business collaborations, industry voice). Interviewees that were able to benefit from networks cited themselves as 'networking personalities' and highlight the proactive nature of their involvement.

The added value of being a member of NEPIC that was cited most was engagement with the cluster organization itself. NEPIC provides a 'trusted platform' for organizations to share information, some of which may be commercially sensitive, that enables NEPIC employees to identify possible collaboration opportunities between members or opportunities for individual businesses. The expertise and sector experience of NEPIC employees was repeatedly cited as important for understanding and tailoring opportunities to members.

Business Opportunities – Integration

NEPIC works with industry leaders to undertake specific projects as a collective group of organizations ('thrust teams'). These projects include: re-integration; up-skilling the local labour market; and specific growth activities. Broadly, these projects aim to contribute to long term sector-level issues for existing but fragmented industries. Interview participants representing multinational organizations with manufacturing sites in the area have cited involvement with thrust teams as the main area of engagement with NEPIC (Member-Multinational-Manufacturer 1; Member-Multinational-Manufacturer 2). In both cases, the activities are industry-led but NEPIC provides support with coordination and management of activities. Co-location advantages from physical integration (e.g. access to feedstock's and waste - including energy - outlets) were cited as important for these local sites, although the benefits of coordination are not necessarily recognized by the (international) owners of the organizations. These activities represent long term investments in wider infrastructure, which would be unlikely to be invested by individual organizations that contribute to the competitive advantage of locating in the area.

These findings are limited due to the sample size, however, the results do indicate difference between the value added for organizations that are part of the core group of 6/7 organizations (integration) and those in the wider network (networking). The core group represents a small, but critical, group of the concentration that has some degree of interaction between firms. The wider network represents a large group of firms in the concentration but the interaction is more piecemeal and not based on co-location benefits.

Table 19: NEPIC business opportunities

Interviewee	Nature of Business Opportunity
Member-SME-Manufacturer 1	A strategic group was established by the participant to rebrand and launch a product. Through a 'Meet the Members' event the participant accessed related professional business service firms (marketing, specialist animation professional from a local university and UKTI). Further specialists were brought in to support the project (language specialist, commerce specialist) that were identified through UKTI. The group successfully launched the product for export markets.
Non-member-SME-Energy	Through the BASME program the participant was able to establish contact with a key potential customer. This has yet to develop into a significant commercial relationship.
Member-SME-Business Services 1	Through the BASME program the participant established contact with eight organizations (for direct and indirect commercial relationships). Opportunity to also build 'kudos' in the market place.
Member-SME-Business Services 2	Participated in the trade mission to India run through NEPIC. This would not have been undertaken independently by the participant and the company had no existing trading relationships with India. No direct commercial relationships have been established yet from this, although it is an ongoing area of potential

Interviewee	Nature of Business Opportunity
	development.
Member -SME-Business Services 3	No business opportunities have been established through NEPIC membership. Participant has been more heavily involved with another networking organization that has generated 50% of business referrals in the first year.
Member-SME-Energy	Participant has existing knowledge of the region and actors and therefore does not use NEPIC to identify business opportunities.

3.2 Financial Services Cluster within Leeds City Region

3.2.1 Leeds Financial: Historical development of cluster

Leeds City Region is the UK's second centre for banking and comprises a significant professional services hub. Leeds is the capital of the wider Yorkshire, a region with a long heritage in financial and professional services, which today combines a significant number of firms of varying sizes. The region is host to national and international well-known companies including HSBC, Santander, Royal Bank of Scotland, TD Direct Investing, Lloyds Banking Group, KPMG, EY, Deloitte and PwC (UKTI 2016). Leeds City Region is also recognised as the home of the building society with three of the five largest UK building societies headquartered in the region. Yorkshire, Skipton and Leeds building societies all have a significant workforce in the region as does Nationwide, the world's largest building society. A large and diverse range of insurance services also operate in the region, from global insurers, customer services and claims handling centres, to underwriting and brokerage operating across various markets and specialisms. Over 13,000 people are employed in banking and over 129,000 in the overall financial and business services sector (ONS 2014) in the region.

The Financial services industry benefits from the region's excellent digital connectivity through its independent internet exchange and has access to the graduates of eleven universities. The area produces over 40,000 graduates in business and associated subjects (HESA data 2013/14) and efforts have been made to develop an employer-led approach to producing business-ready graduates with relevant professional skills. Examples of this include Accounting and Finance, and Data Analytics at Leeds University Business School (LUBS) and two upcoming Financial Exchange Trainee Trading Hubs at Leeds Beckett. In addition a number of non-academic routes have been developed such as the Leeds-Legal Apprenticeship Scheme which provides lectures and mentoring from key business leaders. The City Region is also the home of the Dotforge, the first Financial Services Technology Accelerator outside of London.

3.2.2 Leeds Financial Services: Case study methodology

A desk based scoping exercise was undertaken to develop an understanding of the financial services industry in the Leeds City Region using national and regional reports from statutory bodies and academic literature. The Financial Conduct Authority register

was then used to identify financial services firms based within Leeds by postcode. The register was sieved to include only firms which:

a) Had financial services related core business

b) Had a named individual contact on FCA website with personal email address

c) Had a website

Emails were sent to the sample of firms that met the above criteria (50 plus firms) with follow up emails until a response was achieved. During initial interviews snowballing techniques were also used whereby respondents were asked to identify other potential interviewees. Nine interviews with representatives from firms sampled against size proportional to the sector as a whole were undertaken. The firms cover a range of services including Banking, Credit Checking, and Financial Management. Larger firms may be either regional offices or sub-firms of larger companies. However three of the larger firms interviewed are locally based firms and one used to be a Leeds based firm before it was purchased by an outside investor. Most firms are less than 10 years old although some are much older.

The nature of the sector means that some of the interviewees held, in addition to their private sector roles, roles within public sector or trade organisations. Of these: Two interviews were with private sector figures who had senior roles in relevant Local Economic Partnership (LEPs) and two interviews were with private sector figures who had senior roles in trade organisations. Two interviews were also undertaken with senior officers from Leeds City Council and from the Leeds LEP (Table 20). The researcher also attended the Yorkshire Financial Centre of Excellence Launch where speeches were given by senior government figures, one of which was at Ministerial level, and which gave further insights into the Financial Services industry. The Launch event was also used to identify additional respondents.

Table 20: Summary of interviewee sample

Organisations interviewed	Number of interviews	Position of interviewees	General information about company interviewed
Large multi-national company	1	Head of compliance monitoring	Company has overall turnover in excess of $11billion and over 72,000 employees.
Large UK company	1	HR Director	Company has turnover in excess of £200million (2015) and 1,300 employees
Medium sized companies	4	Chief Officers / Directors Some of these interviewees had additional senior roles in: Leeds LEP Sheffield LEP	Companies with turnover between £10million and £30million.

Organisations interviewed	Number of interviews	Position of interviewees	General information about company interviewed
		Personal Finance Society	
Small businesses	3	Owner operators	Companies with less than 5 employees
Local Authority	1	Head of innovation and sectoral development	
Leeds City Region LEP	1	Head of Inward Investment Operations	

3.2.3 Leeds Financial Services Findings

Origin of the cluster

The Larger firms interviewed tended to be based in Leeds either due to historical reasons – Leeds has been the financial centre of Yorkshire for over 250 years – or because the owners/founders were already based in Leeds prior to a buy-out from a company from outside the area. For the smaller firms interviewed the existence of larger companies in the region had played a significant role in their location within Leeds as the larger companies provided a training ground for people who later went on to establish their own firms. All respondents were happy in Leeds due to its reputation as a financial services centre and the readily availability of customers. Even larger firms report that their main customers particularly like the fact they are in Leeds. It was the belief of many respondents that the role of Leeds as a financial services centre stemmed from the growth of non-conformist mutual societies supporting the mills and factory workers during the industrial revolution. In fact evidence shows that local banks supporting the developing cloth trade were perhaps more important but the notion of the building society and the community values which underpin them were extremely prevalent.

Firm consciousness of being part of a cluster

All respondents are aware of the existence of a ‘cluster’ and appreciate it. The larger firms don’t directly benefit from it too much but smaller firms do, especially through the chance to network with other companies. This is particularly used by firms to share detailed knowledge of emerging changes to regulation or legislation. Such changes are many and varied with a large number of changes to tax legislation for example being difficult for smaller firms to stay abreast of. These networking opportunities enable companies to share best practice around potential investments in order to maximise return for their customers. This is almost exclusively a knowledge exchange practice and there is little to no reported shared product development (indeed in some cases such activity would be illegal in this sector).

For example the interviewees from some of the large firms in the study said:

“The accountancy practices know each other very well in a semi social situation but don’t co-operate in any formal sense. What you do find is the offices of PwC, Grant Thornton,

KPMG, Deloitte etc up here are generally about attracting new clients and the operational issues that go with that. While the senior people spend time in London, the real strategic direction and drivers tend to be driven from London, and the key back office functions are definitely run from London" (Personal Finance Society).

"Within the industry we are aware of Leeds position as the second financial centre of the UK but I'm not sure the average person in Leeds knows that" (Large UK Company).

"The capability here is pretty much the best of any city outside London" (Multinational Company).

It is interesting to note that many firms initially identified by the methodology were small financial advisory services working from home but with a Leeds front office address. Many websites of firms within the relevant SIC codes therefore place these firms within Leeds while in fact the owner/operators may live across Yorkshire. These firms regard having a central Leeds address as important to customers' perception of their business.

"Our business is very much face to face trusted advisor relationships and Yorkshire people like to work with Yorkshire people – a big part of the cluster is that people like to work with local firms" (Small Fund Manager).

Linkages with other firm in cluster

Virtually all interviewees initially stated that they had limited or no links to other firms in the region. However during the interviews it often became clear that this wasn't the case and many interviewees changed their position. This is because the links between businesses are extremely informal, often only existing socially, or through activities additional to work such as trade organisations or societies. Traditionally for example larger firms would organise social events from golfing days to cruises for smaller fund managers. These seem to be less prevalent than they once were and increasingly events tend to take the form of knowledge sharing over dinner or via speeches.

"If you take the Building Society Sector it is actually well known for being quite collegiate. We have a trade association and actually there's quite a lot of discussion and debate and help given to the smaller building societies. It's not anti-competitive but the Chief Execs get together quite a lot to help each other with regulation, maybe lobbying on behalf of each other or developing guidance notes for each other on shared issues such as regulatory change. It doesn't happen below Chief Exec level though" (Large UK Company).

"The cluster in Leeds is based on individuals who run firms rather than the activities of the firms themselves. It's their networks that make the cluster. It's more of an informal cluster in that sense" (Leeds City Council).

An additional way that firms link to each other is the movement of staff from one firm to another. This can happen at a number of levels. At entry level most new starters within the cluster seem to do so in large firms. Once they have developed relevant skill sets often through promotion within specific wings of the company they become more attractive to smaller firms or potentially start up their own companies. Very highly sought skills, such as actuarial risk assessors, tend to be recruited from outside the cluster. While there are a number of trade societies, there is no specific organisation which supports the cluster as an entity and as such the cluster seems to operate in two informal ways:

1. The large firms in the region train staff who then go on to establish their own, mostly smaller firms.
2. It provides networking and knowledge exchange especially around new and emerging legislation normally through trade organisations as discussed above.

"I did 23 years for Lloyds Bank in Leeds straight from school and learned all aspects of branch banking and then relationship management. The last roles were 'hunter' roles which were about going out and introducing people to our products and that exposed me to the broker market. How to deal with clients and approach them is learned in large banks before I set up my own company" (Small Fund Manager).

"I started at the 'Pru' like we all did before we set ourselves up as independent" (Small Fund Manager).

"It's very hard to show in the data but what we are seeing anecdotally is people leaving big firms and starting their own companies not when they are young but when they are older and have built up skills and a network around them" (Leeds Council).

A relevant finding in the context of Big Data methodology is that the websites of the financial firms studied do not highlight the networks in which they operate. Indeed the websites are almost entirely customer focussed and are designed to sell products or, for the larger firms, to allow customers to engage in online banking or financial management. There is little discussion of the actual operation of the firms.

Linkages with firms outside the cluster

All respondents reported linkages across the UK but there were no 'strategically important' areas mentioned by a significant number of respondents. Larger firms have other offices across the UK or the rest of the world and have varying degrees of autonomy from these depending on company structure, personnel, events etc. As in all cases this is likely to be fluid within and between companies often depending upon personalities and changing corporate approaches to regional autonomy. Smaller firms operate almost exclusively within the Yorkshire area (often defined as a vague area from south of Newcastle to north of Nottingham). Other than knowledge sharing as described above – i.e. the detailing of changes to regulation and legislation via networking events they do not work directly with other companies.

The larger firms are aware of skilled personnel in other companies, particularly in London, and will directly head hunt them when possible. Often the fact that Leeds is seen as having a better quality of life is used to support this but for some high demand/high wage staff, especially actuarial risk experts, the large firms accept that they will have to allow them to live in London and commute to Leeds.

Linkages with institutions

There are a number of professional societies which host networking and knowledge exchange events. Some companies do this too either as a mechanism to develop networking with other companies (although rarely is this networking linked to specific outcomes; rather it is seen as a worthwhile activity in its own right) or encouraged by public sector bodies. All respondents reported being active users of such events although it is possible the methodology is more likely to draw out for interview people who are more likely to come forward and that such individuals may be more likely to attend events.

The role of the LEP and of public sector bodies was often mentioned here and almost all respondents spoke well of the activities of the LEP and were supportive of their efforts. In general though few could be specific about these activities and interviewees from the LEPs were open about the limitations of their actions upon the cluster. They pointed out for example that the primary issues affecting businesses within the cluster were either national or international regulation or market shifts which could not be controlled locally. Where they did feel they could add value was in the development of infrastructure and branding to support cluster growth, specifically the development of wider perceptions of Leeds as a Financial Services City to businesses internationally and the housing and amenities which would attract relevant workers – for example more inner city luxury flats. Specific points were also raised about competition between LEP areas within Yorkshire around inward investment. It was felt that teams which should be working together spent a "whole lot of time" trying to persuade companies to jump from one side of a political boundary to another.

Some, especially amongst the larger firms, supported the view that the key issues which affected their business were national or international and there was little the LEP could do in practice to support them. This did not negate the generally positive feeling they expressed towards the LEP and their attempts to promote the region. Two respondents are private sector figures who play a significant role in either the Leeds LEP or neighbouring ones and they were extremely keen to point out the importance the LEP has in promoting the area as the UK's second banking city to potential investors in the area. During the networking event referred to previously it was made clear over conversation that many private sector figures within the industry saw supporting the sector as almost a social duty of their firm.

Value added of locating in a cluster

The nature of linkages and interactions amongst Financial Services firms in Leeds can best be categorised as informal and fluid. There are a large number of firms engaged in similar activities and for the larger firms this allows access to skilled workers for some roles. They also utilise the large graduate pool. Smaller firms are keen to utilise the areas' reputation to enhance their own profile and are often managed by staff who have been trained within larger firms. There are a number of trade associations which support networking events based around knowledge sharing and there are strong personal relationships between those who attend these events. However there is no single organisation tasked with developing the cluster and operational linkages can be relatively weak in practice.

3.3 Birmingham Digital-Health cluster

Digital Health, or ehealth, is an emerging industrial sector based on the application of information and communication technologies (ICT) to the provision of healthcare. Following the Deloitte (2015) report for the Office for Life Sciences the present study uses the following classification of activities as ehealth:

- Telehealthcare (telecare and telehealth): support and assistance provided at a distance using ICT and the remote exchange of clinical data between a patient and their clinician.
- mHealth: mobile phone applications relating to health and/or wellbeing and connected wearable devices

- Health analytics: software solutions and analytical capabilities needed to assimilate big data.
- Digitised health systems: digital health information storage and exchange of digitised patient medical records.

A study of digital technology clusters in the UK undertaken by Tech City UK (2016) identified Birmingham as an important cluster with particular strengths in the areas of digital advertising and marketing, enterprise software and cloud computing, online gambling, and telecommunications and networking. Birmingham's history as a centre of advanced manufacturing was also noted as one of the city's strengths. The Tech City Study identified Birmingham as one of the top five digital employers in the UK and one of the cities where business are most likely to recruit their skilled workers from one of the three local universities (University of Birmingham, Aston University and Birmingham City University). Surveys of digital firms have identified access to affordable commercial property and local networks as attractive features of the city for their businesses. Innovation Birmingham based at the Digital Innovation Campus, a facility owned by Birmingham City Council, will soon launch an expanded facility- iCentrum- to host and support digital start-up firms in the city.

3.3.1 Case Study Methodology

Choice of firms

Twenty-seven firms were initially identified as digital health businesses in the Birmingham area. Of these twenty-seven, seventeen firms were identified by the Office for Life Sciences (OLS), nine firms as a result of the Big Data analysis, and one firm was identified by both sources. Telephone and web-based research was undertaken on the twenty-seven firms to gather more detailed information about the nature of their activities and their location. Of the ten firms identified by Big Data, six firms could potentially be classified as either producers or users of digital health or software technologies, or as management consultancy firms working with these technologies. Of the eighteen firms identified by the OLS, six firms could be identified as producers of digital health or software products sold to the health sector and also had a Birmingham postcode. The twelve remaining firms identified by OLS did not produce digital health or software products sold to the healthcare sector, did not have a Birmingham postcode, had been dissolved, or it was not possible to gain any information about them.

Of the total twenty-seven firms identified, eight firms were confirmed producers of products based on digital health or software technologies which they sold to the healthcare industry, had a company website and were located in the Birmingham area. All eight firms were interviewed. A ninth firm which was identified neither by OLS nor by Big Data was also interviewed. This firm was identified by interviewees from the Queen Elizabeth Hospital as a Birmingham based firm with whom they had collaborated to create a digital health device. A total of nine firms ranging from 1 to 80 employees were interviewed (see Table 21). Of these nine firms four described themselves as digital health and five as software firms.

Choice of organisations

A scoping exercise identified a number of regional organisations relevant to the development of a digital health industry including:

The West Midlands Academic Health Science Network (WMAHSN)

The WMAHSN is one of 15 designated Academic Health Science Networks in England established by NHS England following the publication of Innovation Health and Wealth (2011) report. The stated aim of the WMAHSN is to "lead, catalyse and drive co-operation, collaboration and productivity between academia, industry, health and care providers and commissioners, and citizens, and accelerate the adoption of innovation to generate continuous improvement in the region's health and wealth" (http://www.wmahsn.org/about-us/). Members of the WMAHSN include NHS commissioners, providers of NHS services, industry, academia, the third sector, patients and carers, and the public.

The Innovation Birmingham Campus and iCentrum

The Innovation Birmingham Campus is a digital incubation centre which in March 2016 is due to inaugurate the £8m iCentrum® building as a digital hub for the West Midlands area. iCentrum has received funding from and will work in partnership with the NHS-funded WMAHSN. One of the aims of Innovation Birmingham and iCentrum® is to host digital health companies, offer desk space, encourage networking activities and enable start-ups to access the region's network of NHS trusts and Clinical Commissioning Groups.

The Queen Elizabeth Hospital and the Institute for Translational Medicine

The **Queen Elizabeth Hospital** is another important institution in the healthcare ecosystem of the Birmingham area. The hospital is both a producer and a buyer of digital health products. Linked to the hospital and under the leadership of Birmingham Health Partners is **The Institute for Translational Medicine** (ITM), a new clinical research facility opened in 2015, which aims to facilitate collaborative relationships with industry for the efficient evaluation of new treatments and innovations in NHS practice. The ITM is part of the WMAHSN.

Medilink West Midlands

Medilink West Midlands (part of Medilink UK) is a private membership-based organisation with a remit to stimulate growth and innovation in the health technology sector. Medilink brings together the NHS, academia and businesses and the aim of the network is to provide specialist consultancy services in the areas of innovation and commercialisation, international trade, public relations, marketing and skills.

Interviews took place with the Innovation Birmingham Campus, the Queen Elizabeth Hospital, the Institute for Translational Medicine and with Medilink West Midland.

Table 21: List of companies producing Digital-Health products identified by OLS, Big Data and interviews

Companies	How companies they were identified (web scraping/OLS list)	Birmingham firms identified as producers of software or digital technology products sold to the healthcare industry (after further telephone and web-based research)	Firms defined themselves as digital health in interview	Interviewed
Digital Life Sciences Ltd	BD	✓	✓	✓
Care Monitoring 2000 Limited	BD and OLS	✓	x	✓
Iuvo Limited	BD	✓	x	✓
Oral Health Innovations	OLS	✓	✓	✓
Achiever Software	OLS	✓	x	✓
Safe Patients Systems Ltd	OLS	✓	✓	✓
Inventor-e Ltd	OLS	✓	x	✓
Ccbt Ltd	OLS	✓	✓	✓
Stormnet Media	Identified in interviews	✓	x	✓
Quest Healthcare Limited	BD	x		x
Fontus Health ltd	BD	x		x
Better as.one Limited	BD	x		x
Hampton Knight Limited	BD	x		x
Pharmaspec Limited	BD	x		x
Pharmfirst Consultancy Limited	BD	x		x
Eurofins Agroscience Services Limited	BD	x		x

Companies	How companies they were identified (web scraping/OLS list)	Birmingham firms identified as producers of software or digital technology products sold to the healthcare industry (after further telephone and web-based research)	Firms defined themselves as digital health in interview	Interviewed
Connected Health Consulting Ltd	OLS	x		x
Sero Solutions Ltd	OLS	x		x
Tynetec Limited	OLS	x		x
Alert Life Sciences Computing UK	OLS	Not possible to verify		x
Civica Group Limited	OLS	Not possible to verify		x
Marlbrook (UK) Ltd	OLS	Not possible to verify		x
Cranlea Medical Limited	OLS	x		x
Ahc Information Services Ltd	OLS	x		x
Halliday James Ltd (West Midlands)	OLS	x		x
Just Checking Ltd	OLS	x		x
Lanner Group Ltd	OLS	x		x
Emerson Network Power	OLS	x		x

Table 22: List of all interviews

Digital Life Sciences ltd
Care monitoring 2000 limited
Iuvo limited
Oral Health Innovations
Achiever Software
Save Patients First Ltd
Inventor-e Ltd
Ccbt Ltd
Stormnet Media
Queen Elizabeth Hospital Institute for Translational Medicine
Innovation Birmingham
Medilink West Midlands

A review of the websites of the companies studied shows that firms tend to display information about their main clients, research partnerships with universities and the names of prestigious institutions from whom they have won awards. Public sector organisations such as the NHS, universities and local councils do feature in the websites as they are important clients. Firms display less information, if any, about their supply chains though one of the firms interviewed does display a list of well-known names in the telecommunications and electronics industry- such as BT, Vodafone, Samsung and O2- as partners.

3.3.1 Digital-Health findings

The firms interviewed date from the late 1990s or 2000s and are classified as SMEs. All nine firms are located in Birmingham for fortuitous reasons linked to the original location of the founders at the time the companies were established. The companies interviewed have adopted different business models but the four firms that defined themselves as digital health developed their products in collaboration with the NHS or university medical research facilities and maintain close working relations with consultants or GP surgeries working for the NHS or academic research centres.

Eight of the nine companies interviewed were UK-owned and two of the nine companies were part of a larger group. Two of the four digital health companies studied had significant US investments. One of these companies (Company E) – the only non-UK company in this study – had started out as a UK-owned firm and, with the support of US investor, had grown until 2012 employing up to 25 employees. The restructuring of the NHS during 2012-2015 had forced the company to downsize but it was saved by US investors who acquired the company after licensing its product to a large US insurance firm. Today product development and the writing of software is organised from the US

(though some of this work is probably outsourced to lower costs locations such as India) and the role of the UK group is sales and marketing outside the US. Apart from company E (now US-owned) only one other company (company F) had a facility outside the UK. This was in India where the company had a small in-house team tasked with product development. Another of the firms interviewed had experimented with locating software production in India but had brought this work back to the UK because of quality problems.

Main markets

Except for one firm which has a large US market (Company E) the firms in this study mainly sell their products in the UK. Many of these firms have clients in the public sector which are distributed throughout the UK. The local market is not particularly important for any of the firms considered in this study (see Table 24 for a breakdown of main markets).

Origin of the cluster

We did not find evidence of the existence of a digital health cluster or of a significant agglomeration of digital health firms in the Birmingham area. None of our interviewees thought there was a *Birmingham* Digital-Health cluster although regional bodies such as Medilink and a few of the firms interviewed did identify a *West Midlands* cluster.

Our interviewees identified two main advantages of being located in the Birmingham area. The first is Birmingham's central location in the UK with good quality road and rail networks and a well-connected local airport. This was acknowledged as particularly important as all firms have customers located throughout the UK. The second advantage identified by a number of interviewees is the existence of wealthy and attractive neighbourhoods with good facilities that make living in Birmingham attractive for key workers.

Linkages with other firm in cluster: Importance of links and types of linkages

The importance of linkages with local firms in the Birmingham area differs amongst the companies interviewed. A number of companies do not have any links with other local businesses, services providers or institutions and interviewees did not see any benefits from local linkages. Other companies do acknowledge benefits that arise from local relationships even when their strategic partnerships are outside the Birmingham area. Company A, for example, has partnerships throughout the UK. When searching for a new partner the firm does not start with a local search but uses Google searches. However, if a suitable local firm is identified for partnership they will be approached first as it makes the interaction easier (Company A). Other firms interviewed use a similar approach.

Most of the firms interviewed have internalised their relatively short supply chains with tasks undertaken in the Birmingham area. Those firms that outsource important tasks such as software development agreed that suppliers of software services did not need to be local. One of the firms interviewed for example outsources much of its software development to a software house based in Welshpool (about one hour away) and their electronic expert is located in the south of the UK. The company indicated that in the area of software, suppliers do not have to be local as long as they are not too far away.

There were two firms were interviewees did think that location close to their suppliers was important. The first firm uses specially designed cabinets for their products and needs to be located in relative close proximity (40 minutes) to the supplier responsible for designing and fabricating the cabinets. The second case was the relationship established between the Queen Elizabeth Hospital and Stormnet Media to jointly develop an application for mobile phones that allows the transfer of images. The development of the application required sensitive data to be transferred to Stormnet Media which required the firm to gain the full trust of the hospital's IT department. The development of a high trust relationship required both teams to work closely together for a period of time and the interviewee noted that the strong bond established was unlikely to have developed if the partners had not be located in close proximity. The partnership has been very successful for Stormnet Media as it has enabled the company to enter the health market and develop a new business model based on collaboration with the hospital.

"There are a number of benefits to having a business model based on working with the NHS. We get their backing and their name on the product which gives you a presence in the market from the word go…. [It] also allows you to trial and test the product and gets into all the trials that are needed to make sure the product is robust to meet the needs of the market place and for us that is a key ingredient." (Stormnet Media)

All interviewees agreed that at present there is sufficient high-quality knowledge and expertise in the Birmingham area to service their businesses. One of the larger firms noted however that there is high competition from large companies such as Jaguar for this expertise and there is concern about their ability to recruit the people with the necessary skills at a price they could afford as the company grew.

"There is a high quantity of knowledge-based people in this area. The problem is that there is a lot of competition for them so my worry is that as we grow there will be problems recruiting the right people for the amounts of money that we can afford. As an SME we are competing with Jaguar for an engineer for example." (Company C)

Another company referred to the existence of high-quality skills in Birmingham yet they were aware that the area could benefit from a greater technological profile

"We have never struggled recruiting the people we need but we are aware that on the development side we would get more skills if we were to locate on the M4 corridor because that area is known as a hub of technology. So Birmingham could benefit by having a greater technology profile" (Company G).

For many of these firms however their strategic partnerships are outside the Birmingham area.

Linkages with institutions

The interviews showed differences between firms both with regards to their linkages with local education or research institutions, and with local and regional network organisations.

Two of the firms interviewed have research linkages with local universities- one with the University of Birmingham and one with City University – but these are not seen as strategic to the firms. A number of firms had recently begun to take on apprentices from

local academies and placement students from local universities, but the experience was too recent for firms to be able to evaluate the success of these initiatives.

Firms also differed in their evaluation of the importance of linkages with regional industry networking organisations. Some of the companies interviewed do not see any value in local networking as expressed by company G

"We don't tend to do the whole socializing and network thing with our competitors; we have never seen the point of that. We are busy out there hunting for new customers so we tend not to do the showy networking stuff that some companies may do" (Company G).

Other firms interviewed however are aware of the benefits of linking with local networking institutions. A number of firms had in the past received small amounts of funding in the form of grants from local bodies which had proved important for the development of their businesses. Moreover the existence of the Innovation Birmingham Campus was identified as important by two of the firms interviewed. One of the companies is actually located in the campus and this has enabled the firm to make contact with other electronic businesses in the building as well as a lawyer that specialised in IT-related contracts. The second company uses the facilities of the Innovation Campus and has accessed a number of local services they would not have used had they not been in contact with Innovation Birmingham. The company is aware of the existence of a group of digital start-ups located in the Innovation Campus although at present there is little interaction with them. The firm also attends talks organised by Innovation Birmingham which, while they are aimed mainly at the start-ups based in the building, can sometimes be of interest to them. Interaction of this nature, however, was not the norm amongst the rest of the firms interviewed.

A number of the firms established as digital health business from the beginning- rather than software firms selling products to the healthcare sector- noted the importance of local networking organisations such as Medilink for their development.

"One of things that helped us in the early stages with funding was that we worked on a couple of projects with Medilink. This was very useful both in terms of funding but also the connections we made with a few likeminded start-ups. They were not digital but they were technology oriented businesses. Those types of support services were useful in the early stages" (Company I).

A number of firms however noted the fragmented nature of local networking organizations, and argued that they needed to be more 'joined-up' because the broad nature of digital health required the development of a variety of skill-sets.

Value of being in cluster

In the case of digital health our case study has found no evidence of the existence of a Birmingham cluster, nor of a significant agglomeration of digital health firms in the Birmingham area. The digital health and software firms interviewed tend to have short supply chains which are mainly internal and clients tend to be spread throughout the UK. When partnerships are established location does not appear to be an important criteria influencing the selection of partners. Digital Health and software firms in the region benefit from the supply of skilled labour that is consistent with a large conurbation such as Birmingham but the opinion of interviewees is that the same skills can be found in other

large cities of the UK. The main benefit of locating in Birmingham is its central location within the UK and the existence of a very good transport network. The role of the Queen Elizabeth Hospital as both a producer and buyer of digital health products could become an important factor in the development of a digital cluster in the future but at the moment this is underdeveloped.

Table 23 - Results from company interviews

	Data of establishment	Number of employees	Describes itself as Digital Health	How the company describes itself
Digital Life Sciences ltd	2013	50 employees	✓	Company delivers a combination of digital products along with consultancy services that lead to the physical transformation of work. They mainly work with primary care organisations in the NHS such as GP surgeries to help them access patient and support people with long term conditions using digital technologies. SIC Code: 62012 - Business and domestic software development
Care monitoring 2000 limited	1999	80 employees	X	They provide telephony-based software. The product is a time attendance system that is used in the care sector using telephones to log time spent with service users and patients. Their main market is community care but as healthcare. Is growing in the community they are expanding into the health care market. SIC Code: 62012 - Business and domestic software development SIC Code: 62090 - Other information technology service activities
Iuvo limited	1998	8 employees. Directors and support people	X	They are a middleware provider. They provide electronic communication services to convert data from one clinical system to another using the NHS or medically approved messaging standards. They sit in between different clinical systems to enable the transfer of information securely and efficiently. SIC Code 62020 - Information technology consultancy activities
Oral Health Innovations	2009	1 employee	✓	Sells software to dentists. Dentists use it for assessing whether their patients will get one of the 4 oral diseases. SIC Code 63990 - Other information service activities (chosen randomly)

	Data of establishment	Number of employees	Describes itself as Digital Health	How the company describes itself
Achiever Software	Mid 1990s. A spin-off from a former US company	27 employees	X	Describes itself as a generic software company. Company Categorized under Computer Support Services. SIC Code: 7379 - Computer Related Services
Inventor-e Ltd	2001	6 employees	X	Describes itself as a manufacturing company and an industrial vending and app developer. Their products are related to the management of personal protection equipment in large industrial sites. They see themselves as a cross between a manufacturing business and an IT service provider. The company has Intellectual property. They do not describe themselves as ehealth but would like to move into the health market. SIC Code 1: 28990 - Manufacture of other special-purpose machinery. SIC Code 2: 62090 - Other information technology service activities
Ccbt Ltd	Company founded in the late 1990s. The trading company CCBT was set up in 2004	4 to 5 in UK. Mainly sales and marketing	✓	Company produces on-line mental health treatments. SIC Code 1: 86900 - Other human health activities
Safe Patients First	Company established 2008	32 employees	✓	Company defines itself as digital health with focus on telehealth solutions. Company established by a consultant surgeon working in the NHS. They focus on software-enabled solutions for health care. Company has been embedded in NHS environment since its foundation and co-creates and designs product with the NHS. The company describes itself as clinically founded and driven.

	Data of establishment	Number of employees	Describes itself as Digital Health	How the company describes itself
				SIC Code 72190 - Other research and experimental development on natural sciences and engineering
Stormnet Media	Company incorporated in 2008	7 full-time employees but also use 7 freelancers on regular basis	X Company is now taking steps to move into ehealth market after successful collaboration with QEH	Company started as a video production firm and developed skills in digital imaging. In collaboration with QEH the company produced a digital health app which will sell in collaboration with hospital. As a result the company is considering becoming a digital health business. SIC Code 59112 - Video production activities

Table 24: Main markets of the companies interviewed

Company	Mainly in UK	Export markets
A	UK public sector is main market. Firm also has a tissue and sample tracking product which it sells to hospital laboratories, universities and biobanks	Some sales to a German University
B	UK NHS and independent health providers that provide services to the NHS	
C	Mainly UK	Recently started exporting to Germany, Switzerland and Austria. Will soon start exporting t US.
D	License their software from the US but have developed this for the UK market.	
E		US is now main market
F	UK Health market	
G	Midlands, London, south of England	
H	Throughout the UK	
I	Mainly NHS	Plans to expand to US, Europe and Middle East

Table 25: Outsourcing and partnerships

Company id	Activities in-house or outsourced	Partnerships	Location
A	Company does everything in-house	Their strategic partnerships (e.g. when they need to integrate a new functionality to their products) are spread throughout the country	Recognises that there are some benefits to partnering with local businesses
B	Develop software in-house in the Birmingham area. Hardware is procured via internet		
C	Some activities in-house but a high level of outsourcing	For hardware has partnership with local design and manufacturing firm (40 minutes away) Supplier of software development services can be located anywhere as long as it is possible to meet occasionally. This is partly because the software developers they partner	Close location to designer and producer of hardware necessary

Company id	Activities in-house or outsourced	Partnerships	Location
		with are very high quality. At present their software developers are in Welshpool	
D	Product development in-house in collaboration with university	Scientists at the University of Birmingham are doing research on the data that comes out of the product.	Has received financial support from regional bodies
E	Product development is controlled from the US. Main product is software which can be developed anywhere	US establishes partnerships worldwide for product development	Company does see value in local partnering
F	Produce their product in-house. Their coders are located in India		Company does not work closely with any local firm and key employees are located all over the UK
G	Everything done in-house		Sees little value in local partnering
H	Work undertaken in-house and also outsourced	For mobile technology they have partnered with a company based in Finland. This supplier did exactly what the company needed and their location was not a barrier.	Company does not seek local suppliers as location is not a barrier or enhancer.
I	Products developed in-house	Some partnerships and collaborations when their solutions are integrated with other products as this requires the integration of software systems	Partners are in USA and Germany

3.4 The importance of clusters

The qualitative data collected from three significantly different regions and industries reveal a number of common benefits to being located in a cluster. These can be summarized as follows:

- **Improved firm visibility**. Firms located in Industrial clusters can gain greater market visibility as a result of the existence of a critical mass of similar and related businesses in the area. Visibility can be greatly enhanced if the cluster benefits from a strong positive reputation. This factor can be particularly important for SMEs and firms expanding into international markets where lack of visibility is a challenge.

- **Diffusion of knowledge and good practices**. The diffusion of knowledge and good industry practices through both formal and informal networking opportunities can be easier in clusters. This is important for large and small firms, above all those within industries subject to rapid change.

- **Development of trust**. Clusters can facilitate the development of trust between firms that work together because co-location facilitates more frequent face-to-face interaction. This can be particularly important for innovation or when sensitive information has to be shared.

- **Sharing of common resources**. Firms can benefit from the sharing of joint infrastructure, regional concentration of skilled labour, and specialised service providers found in clusters.

Qualitative data also indicates that cluster organisations can play an important role in the development of businesses in the following ways:

- **They act as a 'trusted partner'**. By acting as a trusted partner cluster organisations can facilitate the sharing of confidential information and benchmarking. This can enable the identification of collaborative opportunities and the diffusion of good industry practices.

- **They facilitate networking**. There are numerous benefits to local networking including the diffusion of knowledge and good practice as well as the identification of new business opportunities.

- **They act as a focus for industry expertise**. The industry and regional expertise found in cluster organisations can be the source of informed advice and support for firms, above all SMEs. Based on their deep local and industry knowledge they can identify new opportunities for innovation and business development for local firms.

- **They are a source for long-term strategic leadership**. Cluster organisations can provide industrial leadership within a region. Effective leadership may lead to greater investment, strengthened infrastructure of skills upgrading.

- **They increase the visibility of a region**. This can strengthen the flows of inward investments into a region and support the internationalisation of local firms.

4. Conclusions

This report combined an innovative quantitative methodology based on website data with a traditional qualitative case study approach to investigate the geographical agglomeration and the functional integration of UK companies within three sectors: Digital-Health, Financial-Services and the Processing Industry. On the one hand, the objective of this research was to generate new evidence on UK industrial clusters. On the other hand, it aimed to test the potentials and the limitations of "big-data" techniques applied to the study of this topic.

The quantitative approach reveals both similarities and differences in the patterns of geographical agglomeration across the three sectors. The UK largest urban areas clearly emerge as important agglomeration areas for all three sectors. For example, London, Birmingham and Manchester were consistently identified as the largest sectoral agglomerations. On the contrary smaller urban areas have a different importance across sectors. This evidence suggests that the companies classified as being part of these sectors are attracted to large metropolitan areas by factors that are common to the larger population of UK companies. These factors are likely to include the proximity with larger product and labour markets, and access to strategic tangible and intangible infrastructures within larger cities.

To control for these factors we investigated the geographical agglomeration of companies from the sectors under analysis by using a modified version of the clustering algorithm that penalizes overall concentration of companies from other sectors within the same geographical area. Many of the clusters were still identified by using this more stringent approach. We interpret this result as evidence of positive externalities from the co-location of similar companies within a specific geographical area. Across the three sectors under study, only Oxford and Cambridge emerge as geographical areas where the relative concentration of Digital-Health companies is at least two times greater than the national average.[21] This result points to the influence of very important sector-specific factors in these areas.

The analysis of the network of web-links extracted from companies' webpages reveal interesting differences across sectors. In particular, we find that the websites of Digital-Health and Processing-Industry companies contain frequent links to the websites of academic institutions, and that academic institutions have a very central position in the link-networks of these sectors. The same cannot be said for the Financial Sector where instead links to the same government websites are frequent and common to many companies. Links to companies from the same sector are more frequently found in the Financial Sector and in the Processing Industry than they are in the Digital-Health sector. Although we cannot infer specific relationships between companies and institutions from the analysis of the "link network", this evidence is suggestive of important differences across sectors in terms of functional relationships between companies and institutions.

The qualitative case studies provide interesting information that complement and contrast with some of the findings from the quantitative analysis. First, the case study exploring

[21] Relative density is measured as the ratio of companies from the sector over the total number of companies in that area.

Digital-Health companies in Birmingham reveals that out of ten companies that we classify as Digital-Health based on website data, only six are related to a strict definition of this sector. The other four cases more generally relate to the health industry, in particular to pharmaceutical companies. That same case study suggests that the agglomeration of Digital-Health companies in Birmingham is not generally perceived as a functional cluster, and that there are not significant partnerships between the companies in the area. This qualitative evidence is in line with the limited number of inter-company web-links for this sector. The case studies on the Financial Sector in Leeds and on the NEPIC cluster organisation in the Teesside Valley emphasize the important role of historical legacy and central organisation in the establishment of these clusters. These two case studies shed light on the benefits arising from co-location of companies in the same industry or closely integrated industries.

Interestingly, all case studies confirm the different role played by company-university relationships across sectors that we could infer by comparing the network graphs of web-links. While companies in the NEPIC cluster and in the Digital-Health sector report strategic relationships with universities this is not the case for financial companies in Leeds.

Overall, we conclude that Internet data and data-mining techniques are useful tools to identify the economic activity of companies. These techniques appear more useful when applied to industries with clear characteristics but that cannot be easily classified by SIC codes. This is especially the case for the Digital-Health sector, which has as a distinctive feature the application of digital technologies to human health, but which includes companies operating at the cross-road between different economic activities such as programming, health care and consultancy.

Admittedly, the classification methodology underpinning our quantitative analysis can be improved on a number of dimensions. First, we believe that a supervised learning approach may reduce considerably the amount of noise in the classification stage (i.e., reduce the number of false-positives) and increase the power (i.e., reduce the number of false-negatives) of the classification exercise. This would require constructing a sufficiently large “training” sample where we know *ex-ante* which companies are part of the sector and which ones are not part of it. Based on this training set it would be possible to improve the performance of the algorithm by comparing systematically the algorithm ability to separate firms that truly belong to the sector from the others. The construction of this training set requires industry-specific knowledge and it involves a very time-consuming process of manually identifying a large number of potential false-positives.

One of the objectives of this study was to test the feasibility of a “big-data” approach to industry classification that can be implemented to map geographical clusters of different industries. Our experience reveals that while some general features of the methodology can be similarly applied to different industries, there are a number of industry-specific issues that require ad-hoc solutions. For instance, an initial investment of human work is certainly required to fine-tune the algorithm that identifies a specific industry. However, once this tool is in place it can be used to trace the evolution of particular industries over time in a quasi-automatic way (i.e., with very little human intervention). Therefore, big data methodologies may be particularly useful for updating lists of companies from a specific industry and for tracing the evolution over-time of geographical clusters.

References

M. Anyadike-Danes & K. Bonner & C. Drews & M. Hart, 2013. **Localisation of Industrial Activity across England's LEPs: 2008 & 2012**, ERC Research Paper No.15.

BIS 2014. **BIS Research Strategy** 2014-2015. BIS Research Paper No.1985.

K. Chapman, 2005. **From 'Growth centre' to 'cluster': restructuring, regional development, and the Teesside chemical industry**, Environment and Planning A, vol. 37.

I.M. Clarke, 1985. **The Spatial Organisation of Multinational Corporations**, London & Sydney: Croom Helm, p.287.

M. Delgado & M. E. Porter & S. Stern, 2014. **Defining Clusters of Related Industries**, NBER Working Papers 20375.

Department of Health, 2011, **Innovation Health and Wealth: Accelerating Adoption and Diffusion in the NHS**, https://www.gov.uk/government/news/accelerating-adoption-of-innovation-in-the-nhs

Deloitte, 2015. **Digital Health in the UK An industry study for the Office of Life Sciences,** BIS/15/544 – Digital health industry study: UK market analysis.

G. Duranton & H. G. Overman, 2005. **Testing for Localisation Using Micro-Geographic Data, Review of Economic Studies**, Oxford University Press, vol. 72(4).

G. Duranton & H. G. Overman, 2008. **Exploring The Detailed Location Patterns Of U.K. Manufacturing Industries Using Microgeographic Data**, Journal of Regional Science, Wiley Blackwell, vol. 48(1).

M. Ester, H. Kriegel, J. Sander and X. Xu, 1996. **A density-based algorithm for discovering clusters in large spatial databases with noise**, AAAI Press, Proceedings 226-231.

FAME, 2015. **Company report: Various [online]**. Bureau van Dijk. Available from: https://fame2.bvdep.com (accessed 22/03/16)

C. Ford, 2015. **North East Cluster Group NEPIC Announces £2.5billion GVA Generation [online]**, The Journal, 19:30 7 June 2015. Available at: http://www.chroniclelive.co.uk/business/business-news/north-east-cluster-group-nepic-9399858 (accessed 13/01/16).

S. Higgins, 2013. **Cluster Excellence**, Chemical News, July, 12-14. Available at: http://www.nepic.co.uk/WebformFolder/IndiaChemcialNews_ClusterExcellence_July2013.pdf (accessed 13/01/16).

House of Commons, 2009. **North East Regional Committee - First Report: Industry and Innovation in the North East of England**, Session 2009-10 North-East Regional Committee Publications, Parliamentary, 18 December 2009. Available at: http://www.publications.parliament.uk/pa/cm200910/cmselect/cmneast/169/16902.htm (accessed 13/01/16).

T. Lammer-Gamp, H. Kergel and M. Nerger, 2014. **Cluster organisations in Europe – insights from Bronze and Gold Label Assessments**, Input paper for the workshop "Moving forward the EU policy agenda on cluster excellence", Brussels, September 23rd, 2014. . Available at: http://www.corallia.org/images/CoralliaInput_Paper_COM_Workshop_September_published-2014-09-23.pdf (accessed 13/01/16).

M. Nathan & A. Rosso, 2015. **Mapping digital businesses with Big Data: Some early lessons from the UK**, *Research Policy,* in press.

M. Nathan & E. Vandore, 2014. '**Here Be Startups: Exploring a young digital cluster in Inner East London**' , *Environment and Planning A*, vol.46(10).

NEPIC, 2015a. **NEPIC Directory 2015**, NEPIC. Available at: http://www.nepic.co.uk/wp-content/themes/itchyrobot/directory/Directory2015-16.pdf (accessed 20/01/16).

NEPIC, 2015b. **North East of England Process Industries National Centres of Excellence**, Research Institutions & University Core Specialisms, NEPIC, September 2015.

NEPIC, 2015c. **NEPIC Cluster Strategy: A Strategy to Support Northeast Process Industries, Draft 3** - 17/09/15, PowerPoint presentation provided by NEPIC.Tees Valley Unlimited (n.d.) NEPIC [online]. Available at: https://www.teesvalleyunlimited.gov.uk/tees-valley/invest/key-sectors/chemicals-and-process/nepic.aspx (accessed 13/01/16)

ONS, 2014 **GVA for Local Economic Partnerships 1997-2012.**

M. Porter, 1998. **Clusters and the new economics of competition.** Harvard Business Review, 76(6).

M. Porter, 2000. **Location, Competition and Economic Development: Local Clusters in a Global Economy**, Economic Development Quarterly, vol.14(1).

TechCity UK, 2016. **Tech Nation2016: Transforming UK Industry**. http://www.techcityuk.com/wp-content/uploads/2016/02/Tech-Nation-2016_FINAL-ONLINE1.pdf?utm_content=buffer2e58f&utm_medium=social&utm_source=twitter.com&utm_campaign=buffer

Annex

Interview questionnaire

The primary aim of this research project is to validate different types of methodologies and data such (e.g. Big Data and semantic web methods based on information companies provide on their webpages, econometrics data based on ONS data and interview data) for the mapping of industrial clusters in England. We are also interested in gaining more in-depth understanding about the nature and strategic importance of your collaborations with other firms and stakeholders (e.g. universities, FE colleges and training institutions, R&D consultancies) inside your regional cluster compared to relationships with partners outside the cluster.

The project is being financed by the Department for Business, Innovation and Skills (BIS) and carried out by a team of researchers at Birmingham Business School, University of Birmingham, and the National Institute for Social and Economic Research.

The team at Birmingham Business School has been tasked to do the interview data. If you agree we would like to record the interview to ensure that we have accurate records of the data. When the interview is transcribed we will anonymise your identity (interviewee) and that of your firm (or other organisation) before the data is stored, analysed and shared between researchers or prepared for presentation or publication.

Are you happy to proceed to the interview and would you be happy for us to record the interview. If at any stage you want the recording to stop please let me know

1 Questions about the firm/institution being interviewee and interviewee

<u>About interviewee:</u>

- What's your role in the firm? (formal position)
- How long in the firm?

<u>About the firm:</u>

- Can you tell us what your firm does? What kinds of products/services?
- Identify how the firm classifies itself (**SIC or industry classification**)
- How long has the firm been running?
- Are you part of a bigger (UK/international) company? (*if part of another company would be interested to know something about their position in the value chain*)
- Do you have branches elsewhere in the UK? Overseas?
- How many employees?
- Where are your important markets (in cluster, UK, export)?
- Where are your main (strategically important) supply chain partners
    - in cluster,
    - UK,
    - international?

- Where are your main (in terms of the number of partners) supply chain partners (in cluster, UK, international)?

2. Questions about the origin of the cluster

- How did the firm become located here? (when/history)
- Reasons for location
- has firm ever considered leaving/relocating? If so, what prompted that and why not?

3. Questions about the firm consciousness of being part of a cluster

- Would you consider yourself as part of a cluster? How would the firm define the cluster?
- What type of activities/tasks?
- Reputation of cluster?
- Do you participate in cluster activities (networking events, cluster organisation)?
- Is participating in the cluster relevant to your firm or could you be located anywhere else?

4 Questions about linkages with other firm in cluster: Importance of links and types of linkages

- Explain importance of cluster in terms of supply chain (*and for which part of the supply chain*)?
  - is locating in cluster necessary; important but not necessary; accidental and could be located anywhere else; irrelevant
  - Is locating in cluster important now, in the past, in the future?
  - Examples?
- Do they link up with firms from the same industry or different industries (examples needed)
- Identify key partners within cluster (other firms such as suppliers, customers**) (we need names here so we can interview them)**
- Are your key strategic supply chain relationships with firms inside or outside the cluster/ (examples?)
- Do they find specialised business services in the cluster? (how important are these for the business?)
- How much sub-clustering is there within the sector e.g. do telehealth and mHealth companies in the same area have any connections?

5. Questions about linkages with firms outside the cluster

- Are there other strategic links the company has outside the network?
- How did those links come about?
- Do you have links to any other clusters? Do these relate to your own cluster (e.g. cluster-cluster relationships)

6 Questions about linkages with institutions (non-other firm in cluster): Importance of links and types of linkages

How important is the cluster in terms of supply chain? (1-10)

How important are connections with telehealth and mHealth companies? (1-10)

- Which institutions (other than firms) within the cluster are important to you and why?
- *How important are institutions within the cluster in facilitating knowledge exchange and sharing of best practice?*
- *What is the added-value for firms being in a cluster compared to being located anywhere else in the UK? (skills, knowledge, shared projects etc.)*
- *How important are links with universities and colleges in securing a) the required skills and b) generating new ideas and production methods to increase productivity in the sector?*
- *What makes the cluster function effectively as a cluster, as opposed to a collection of firms that happen to be working in the same sector?*
- *What is the role for Government (local and national) in encouraging cluster development?*
- *What are the barriers to growth for existing clusters?*
- *How important are other sectors to the development of the cluster? E.g. the role of business services…In what way are they important*
- What support do you receive from local government?
- Role of LEP? ….Do you have any contact or support from the LEP?
- Do you have links to institutions nationally (e.g. Innovation Catapults) or internationally?

7 Drawbacks/Problems/challenges of locating in clusters?

- Possible (price of real estate; loss of IP; expensive skilled labour; shortage of skilled or other labour)?

8 How do you see your business developing in the next 5 years?

- Is being part of the cluster important for the development of the business?
- If yes… in what ways?

Big Data and Cluster research: a critical appraisal of the report

by Max Nathan (University of Birmingham)

1 / Big data

'Big data' is generally defined in terms of the Four V's: volume (massive datasets, with millions or billions of observations); velocity (data which may be available at real time or close to it) and variety (a wide range of sources which help us to observe, or model, phenomena previously hard to observe). These features should be useful to policymakers in a range of fields, including business support and local economic development. However, the fourth V is veracity – which throws up a number of challenges for analysts. Raw data is often 'unstructured', and may need substantial cleaning before it's ready to go. Similarly, many commercial datasets are incomplete, but the sampling frame is not always clear (for example, web-scraped data will miss firms without websites, or who have non-scrapable sites). For policymakers, this means taking care with cleaning, validation and interpretation of 'frontier' data sources.

'Big data' comes in three main flavours (Arribas-Bel 2014). These are: data from sensor networks and other sources 'in the wild'; corporate datasets, both internal business data and online sources (from search, social networks or company websites); and administrative datasets, especially microdata: these may be online and open; or available to researchers through resources such as the UK Data Service (UKDS). There are important issues around price and access for some commercial datasets, so it is important to explore the potential of 'public big data' alongside more high-visibility commercial sources (Einav and Levin 2013).

Public debate about 'big data' tends to conflate data sources, datasets and data science techniques. Varian (2014) provides a helpful discussion of the three, focusing on data science tools that can be used for storage and management; diagnostics; and modelling relationships. More broadly, using big data does not mean that theories of change, microfoundations or research design become obsolete, as some enthusiasts have suggested (Anderson 2008). Millions of observations make it easier to find (small) statistically significant associations in the data, which may or may not be meaningful. Similarly, policy based purely on observables will likely lead to a change in observable behaviour in the target population, so that the intervention is ineffective.

2 / Clusters

The idea of industrial clusters has its roots in Alfred Marshall's pioneering work on 'industrial districts' (Marshall 1918) and Jane Jacobs' analysis of ideas-driven urban economic change (Jacobs 1969), as well as more recent work by Hall, Scott, Storper and others on local milieux and 'untraded interdependencies' (Scott 1988, Storper 1997, Hall 2000). The report usefully defines clusters along three dimensions: physical co-location; institutional presence of key actors, such as 'activist' universities and public agencies; and functional relationships between these actors. Beyond this, however, there's little agreement amongst academic analysts about defining and measuring clusters, as well as whether 'cluster policy' is a helpful idea.

Porter (2003) develops these basic features – co-location, institutions, relationships – into the famous 'Porter Diamond', in which local asset bases or 'factor conditions', demand conditions, related industries and 'firm strategy' all positively interact to produce a virtuous cycle of growth. Policymakers should then seek to map clusters on the ground before 'upgrading' them through supply-side interventions. However, Martin and Sunley (2003) suggest that in practice, Porter-style clusters are hard to draw physical boundaries around. Duranton (2011) and Nathan and Overman (2013) point out that clusters also involve feedback loops that are negative for at least some participants (increased competition for workers and market share, higher operating costs), even if these are welfare-positive on the aggregate. This makes cluster policy considerably more complex than Porter's analysis suggests. And in its strong form, the cluster concept assumes that economic linkages between firms and other actors are all within the cluster boundary, whatever this is. In practice, we know that workflows (e.g. supply chains, customer markets) and contextual factors (e.g. national government policy, technological shifts) operate well outside local areas (Bathelt, Malmberg et al. 2004, Saxenian 2006).

Nevertheless, most analysts agree there is some value in trying to descriptively map clusters in terms of one or more of the three dimensions, even if the extent of localised activity and policy action is limited. There are real challenges in observing each in practice: many of these challenges may be amenable to big data and/or data-science driven solutions.

For instance, studies of physical co-location tend to proxy real-world co-location patterns using firm or job shares, or Location Quotients, in standard administrative units (Anyadike-Danes, Bonner et al. 2013). However, such standardised spatial units may not capture actual co-location patterns very well.[22] Duranton and Overman (2005) improve on this by using postcode-level administrative microdata to site firms in continuous space, and develop probabilistic measures of co-location versus a hypothetical randomised allocation of firms across that space.

As this report points out, however, all these contributions are still working with standardised industry codes, and even at a high level of detail (Duranton and Overman use four-digit SICs; five-digit detail is now available) such SICs may not capture emerging economic activities of interest to policymakers. SICs are necessarily backward looking and lag real-world industrial and technological change (Nathan and Rosso 2015). New insights

[22] The Modifiable Unit Area Problem, or MUAP.

from big data may be able to shed light on emerging industrial clusters that current SICs cannot see.

Similarly, there are multiple descriptive studies that 'map' institutional components of clusters such as universities, public-private partnerships, key firms and so on. These act as a useful first layer but often provide little structured information about what connections exist and how these have evolved over time: Garnsey and Heffernan (2005), on the Cambridge tech cluster, is one interesting exception. Data-driven approaches may be able to help with these issues too, although it seems likely that they will need to be complemented by other elements. For example, Nathan and Vandore (2014) combine BSD microdata analysis with interviews to show that the East London tech community evolved from the late 1990s to the early 2010s with little anchor institution or policy involvement.

Functional relationships within and centred on physical clusters are very hard to observe at scale without bespoke data. Notably, Hall and Pain (2006) use origin-destination business phone calls and email info to examine firm-firm linkages across the London city-region. In practice, such internal corporate datasets are still very challenging for researchers to access. Social network data such as Twitter and LinkedIn may offer some promise here; again, quantitative analysis may need to be combined with qualitative components.

3 / Using big data and data science in cluster analysis

To date, big data and data science techniques have been applied in a small number of cluster analyses. In the main, they have been used to improve industry classifications, which have then been used alongside standard area-level measures of spatial co-location. Studies have used both off-the-shelf datasets from a range of sources, and applied data science methods to develop their own metrics.

Catini and colleagues (2015) develop a bibliometric approach to trace cluster boundaries, by using the institutional address fields of researchers publishing in biomedical science journals. Along similar lines, Kerr and Kominers (2015) use inventor addresses from US patents data, exploiting detailed technology field information to draw out a range of cluster shapes which they relate back to industry characteristics and field-specific workflows.

Three recent UK studies have used a combination of open administrative data and unstructured datasets to develop new measures of economic activity in emerging sectors/fields. In their study of the computer games industry, Mateos-Garcia and Bakhshi (2014) use information from online games directories, review sites and industry wikis to develop a detailed list of gaming firms and their locations, which they match to Companies House information. Nathan and Rosso (2015) use web-scraped, modelled sector and product classifications developed by Growth Intelligence on top of Companies House data to provide alternative estimates of UK digital technology firms; extensions to this analysis use BSD microdata, providing high quality information on location, employment and revenue. The most recent Tech Nation report develops a multi-angle take on the tech economy, using Growth Intelligence and Companies House data alongside a number of other unstructured sources including online job ads and meetups. (Using Companies House data presents some problems in identifying physical co-location: more on this below.)

Very few studies have attempted to get a handle on relational aspects of industry clusters. The London and Cambridge Tech Maps[23] use live Twitter data to show mentions and retweets of local firms, although no attempt is made to back out what this activity signifies or its economic importance. Mateos-Garcia and Bakhshi (ibid) highlight a number of suggestive relational and institutional findings for computer games hubs – presence of SIC-related industries, universities offering specialist courses; high levels of residential broadband penetration – but do not try and draw structured connections. Tech Nation (ibid) uses interviews with firms and local agencies to give an impressionistic sense of local ecosystems. Nathan and Rosso are currently working with experimental lifecycle 'events' data from Growth Intelligence, taken from news sites, which provides modelled information on relational activity such as mergers and joint ventures: this may provide the basis for exploring connections within local milieux.

4 / Using qualitative methods alongside 'big data'

Complementing big data-driven approaches with other methods, especially qualitative techniques, is one way to gain a richer understanding of industrial clusters. There is a long tradition of mixed-methods research in economic geography, typically combining aggregate secondary data analysis with questionnaires and/or semi-structured interviews. However, the quantitative element is often simply used to set the scene. Larger, richer microdata offer the potential to develop a more integrated, 'layered' research design (see Nathan and Vandore (2014) for one UK example). These approaches may be especially valuable for research on complex phenomena such as clusters, in which physical, economic and socio-cultural features are all in play. Specifically:

- Qualitative approaches can be used to help a **better understanding of the processes that generate unstructured quantitative data** – for example, understanding an implicit sampling frame in the raw data, or data coverage issues;

- Qualitative techniques may help us **test assumptions in the quantitative analysis** – for example, determining suitable spatial boundaries for a cluster, or understanding the geography of key processes and markets;

- Qualitative methods can **help answer further questions** thrown up by data-driven elements of the analysis – in particular, the extent and nature of relationships between co-located firms and other actors, or the range of business models and customer markets under a broad sectoral label.

In turn, these benefits depend on close linkage between the quantitative and qualitative elements. Specifically, quantitative analysis using large microdata sets can improve the sampling strategy for qualitative analysis – for example, such datasets may provide a more reliable sampling frame than membership organisations or manually-generated lists; if individual identifiers are available, contact and recruitment is easier; large datasets also allow for repeat sampling to improve overall response rates.

[23] http://www.techcitymap.com, http://www.camclustermap.com (accessed 13 April 2016).

5 / The report

5.1 / Summary of methods

The project combines big data-driven quantitative analysis and semi-structured interviews. The quantitative analysis deploys company-level open administrative data from FAME (based on Companies House) with raw information scraped from company websites, and is related to techniques used in Nathan and Rosso (2015) and the 2015 Tech Nation report (Tech City UK 2015). A working sample of companies is developed by removing companies without websites in FAME, and permitting one website per company. 'Restricted samples' of companies for digital health, finance and processing are flagged (from an Office of Life Science list, finance SIC codes and the NEPIC membership list respectively). Scraped data for these businesses is developed into a benchmark set of text 'entities' for each study sector, using feature extraction routines and scored for relevance using TF-IDF type analysis. Other closely related text entities are identified from the remaining data using a word-to-vector (W2V) scoring algorithm. The data is then pooled and each company's entities are W2V-scored against the extended benchmark entity set. By comparing distributions of W2V scores in restricted samples vs. the rest of the data, companies are shortlisted as likely 'digital health', 'finance' or 'processing' if their W2V score is more likely to occur in the relevant restricted sample than in the rest of the sample. For each sector, shortlisted companies with the highest W2V scores above these cut-offs are tagged as 'digital health', 'finance' or 'processing'. Companies House address info is then used to locate the companies. Physical clusters are identified using a modified DBSCAN algorithm: for each study sector, any company must have at least n other companies within *k* kilometres, and this concentration must be greater than the underlying concentration of all firms in that *km* range. Sector-specific firm minima and boundaries are tested and established. Functional relationships are explored using URLs on firm websites.

Following this, the qualitative analysis involves semi-structured interviews with digital health companies in Birmingham, finance businesses in Leeds and processing companies in the North East. Digital health firms are sampled from the quantitative analysis; processing firms are sampled from the NEPIC list used to generate the initial restricted sample and entity set. Finance firms are sampled from the Financial Conduct Authority register, but there is no direct link to the quantitative analysis. A standardised topic guide is used; findings are coded manually.

The project has achieved an impressive amount in a restricted timeframe. The report makes a number of useful contributions: in particular on classifying emerging economic activities, location metrics, and on filling in local detail. The qualitative analysis also helps test some of the propositions / assumptions in the quantitative work. The relational analysis using URLs is less persuasive in its current form.

5.2 / Quantitative methods

I have a number of suggestions for the quantitative methodology, which could be developed in follow-up research.

First, there are some questions around the generation of the starting sample. What are the implications of requiring that the company website be reported in FAME? Who's left out? Is there any patterning to the set of firms whose websites can't be crawled? Ideally one

would want to know if this is a random set or if some types of activity are over-represented. What about firms that provide no SIC information? More broadly, restricting each company to a single website drastically reduces the number of firms in play. If I understand right, this is effectively correcting for corporate legal structure, if we think of websites as a firm's 'front end' and company entities representing (aspects of) the back end. Really one should clean for corporate structure first, then make any adjustment to website allowances that's still necessary. Nathan and Rosso (2015) suggest some ways to do this cleaning, exploiting shareholder information in FAME and using reported revenue.

Second, as the report acknowledges, ideally one would swap out the rule-based classification with a fully machine-learnt approach. To do this, one needs a reliable training set of companies that are known to belong to a given sector or set of sectors. As it stands, the entity set in each restricted sample is only as good as the starting set of firms. For finance, we have an established set of SIC codes representing mainstream financial activities fairly well. For processing and digital health, we have manually curated lists. The qualitative analysis provides some valuable quality testing on these: as I understand it, the NEPIC membership list is quite decent but the OLS digital health list includes at least some companies which can't be understood as 'digital health' even on the broadest framing of the term (this is also evident in the quant analysis, where the researchers end up manually restricting the set of firms in play using SIC codes).

Ideally, one would start with a fine-grained objective vocabulary for each sector – which could then be used to designate a training set, and then to search for the same or closely related terms on websites. Failing that, an agreed set of SIC codes (with tighter and looser variations) might also work (although better for established activities than emerging ones). Developing such a vocabulary / SIC set is obviously challenging for a nexus of activities such as digital health, where understanding boundaries and composition is part of the research question. In this case, an iterative approach may be helpful, in which exploratory qualitative analysis is used to refine understanding and inform the modelling, for example by suggesting key terms around products, services, platforms, customers and business models. Alternatively, patents provide a very detailed set of technology field codes: patent titles and abstract text for a suitable subset of fields could be mined to develop vocabularies for use here.[24]

5.3 / Sectors and sector groupings

A nice feature of the analysis is that it allows for within-sector segmentation that is bespoke for each sector rather than generic. This seems to work particularly well for digital health, which is a hybrid in sector terms, and seems to involve a host of different activities in which digitisation enters in different ways (the qualitative analysis sheds more light on this). SICs alone cannot capture this. Similarly, the data-driven analysis seems to work well for processing, uncovering a range of activities that is rather broader than those suggested by SICs (again, the qualitative analysis tends to confirm this). For financial services, the data-driven approach appears to add less value, with the underlying SICs of the identified firms quite close to the starting set. In turn, this suggests that sector identification based on web / social media information sheds more light on emerging / evolving industrial activities than mature / established ones. This chimes with the earlier discussion about where SICs are more or less 'laggy'. For policymakers, it provides some

[24] Ralf Martin and colleagues at Imperial College London are doing some work along these lines.

useful pointers about where new, big data-driven typologies might best complement existing ones.

These results will also reflect the fact that SICs are self-assigned in the Companies House data that underlies FAME. For emerging activities such as digital health, where no straightforward SIC classification exists, and companies' own business plan may be unclear, we are likely to see more firms assign into generic SIC categories such as 'Other business services'. It would be useful to re-run these results using BSD data, where SIC information is of higher quality.

5.4 / Cluster mapping

The report develops quite a sophisticated cluster mapping routine based on the DBSCAN algorithm. From a researcher's point of view this has some nice features, in particular the flexibility to adjust minimum firm size and spatial boundaries, which allows more naturalistic modelling of specific workflows, labour markets and so on than one-size measures such as Location Quotients (LQs). By using continuous distance measures rather than spatial units, the researchers also avoid boundary/edge issues. However, the new measures are complex and not straightforward to present.

The clustering analysis also involves manual decisions – such as the 'optimal' *eps* and n – which introduce subjective judgement into the analysis. (Why is *eps* never set to 5km or less, for example?) It would be good to find some ways to check the sensitivity of these judgements: qualitative analysis is one way to do this. Conventional LQ-based measures could be provided in an appendix for comparison purposes.

The analysis finds that digital health firms are densely co-located, while firms in finance are less densely clustered; the interpretation is that finance covers a broad / mature set of activities where firms are attracted by urbanisation economies. However, this result is partly driven by the starting definition of 'finance': it would not apply to specialist functions such as investment banking, which are likely to be very densely co-located within a few cities.

The big issue for the cluster mapping is the reliability of the address data in FAME, which is based on registered addresses in Companies House rather than an actual trading address. For younger and/or single plant firms these addresses may be the same or relatively close together. For older and/or multi-plant firms, especially those with a separate HQ function, registered and trading addresses may be quite far apart. This is likely to produce a) misleading cluster mapping and b) an upwards bias in counts for London and larger cities, where HQs are likely to co-locate. There are signs of these problems in the analysis for all three sectors, but they are particularly clear in processing, where on the preferred n and *eps* values, the largest clusters of firms turn up in and around London and Manchester: the North East barely features in firm counts, even though the research starts from the knowledge that there is a large processing cluster in the region. The qualitative analysis provides a useful cross-check on this issue, with a number of interviewees highlighting the HQ problem (and worrying that this ascribes economic output to HQs rather than plants). Using plant-level data such as the BSD would deal with this.

5.4 / Functional relationships

The report has a go at exploring the functional relationships for firms in the study clusters by looking at the counts and types of URLs on their websites (government, companies, other organisations). This kind of thing is extremely hard to do. And on its own, this analysis is going to tell us fairly little: we can't observe why the links are there, how long they have been there, what kind of connections they may represent or their importance. More fundamentally, the set of firms with such 'weblinks' appears to be highly selected (e.g. 80% of digital health firms don't have such links; comparable figures aren't given for other sectors). Given this selectivity, ideally we need to have some idea of firm characteristics for those with and those without weblinks so that we can ascribe some meaning to the results. Directly questioning firms through surveys or interviews can give us a qualitative sense of this. Complementing weblinks with other relational datasets, such as Twitter might also shed some more light on it.

5.5 / Qualitative results

In theory, qualitative approaches can help test and gapfill big data-driven quantitative analysis of industrial clusters. To do this, we need a research design that generates samples for the qualitative element of the analysis directly from the quantitative element. The report manages this for the digital health interviews, where the set of Birmingham companies comes straight out of the results of the data-driven analysis, and to an extent with the processing interviews (as interviewees are taken from the same NEPIC list that forms the restricted sample in the quantitative processing). However, the connection is less clear in the finance interviews, since these are generated from a national FCA register where the overlap with the results of the data work isn't set out. So the financial sector material is interesting in its own right, but can't really be used to interrogate the quantitative elements of the project.

The processing sector interviews throw up a number of useful findings that reinforce / test / complement the quantitative analysis. Interviewees confirm the key finding from the data work that the set of local cluster firm activities is substantially broader than the relevant SICs suggest. They also highlight the firm HQ / location issue discussed earlier in relation to the cluster mapping. Interviewees also provide helpful additional information on the internal structure / history of the processing field in the North East, specifically the divide between ICI and related business in Teeside, and a wider set of firms in the rest of the region. Interviews also highlight the various kinds of proximity in play: not simply geographical co-location, but also organisational proximity (e.g. ICI and related) and the professional communities of interest formalised in NEPIC. It would be useful to extend this line of questioning to see what quantitative / online metrics, if any, might pick up some of these linkages.

Given the quantitative results, the Birmingham digital health interviews have to be seen as representing a collection of co-located firms that has not yet formed into a Porter-style cluster in the city (and may not do so). This highlights that physical co-location does not imply functional relationships; firm location decisions are entirely related to where founders happened to be living at the time. This finding is similar to interview feedback from the first cohorts of tech firms in Silicon Roundabout, before that cluster achieved critical mass (Nathan and Vandore 2014). More broadly, it's notable that the qualitative research also

picks up a West Midlands digital health network, Medilink, which does appear to reflect some cognitive and organisational proximities in the wider region.

The qualitative research also provides some useful supporting information on the diversity of products and business models in play under the digital health banner, and the varying extents to which local upstream and downstream linkages are important to the current crop of firms. This could provide useful information to help segment the sector, as well as providing further inputs to future quantitative analysis.

6 / Summary of recommendations

I conclude with a summary of recommendations for further analysis. These suggestions build directly on what the project team has already done.

- Datasets - matching FAME data to BSD data and re-running the analysis would allow for much more reliable location mapping, as well as analysis on employment and revenue.

- Sampling frame - running some sensitivity checks on sample construction, in particular exploring excluded companies (those without websites on FAME), those removed by the 1:1 company : website condition, and controlling for corporate structure.

- Classification - exchanging the current rule-based classification scheme for a fully machine-learn approach. This is a substantive piece of work in its own right. As the authors note, this requires developing robust training sets of companies for sectors / sector-groups of interest. These might come from membership lists such as NEPIC/Medilist/FCA Register, though these evidently need manual validation first. Other options would be to explore technology field information through mining patent titles / abstracts: this would allow direct identification of applicant firms as well as generating tech-specific vocabularies.

- Cluster mapping - further sensitivity testing on the modified DBSCAN algorithm.

- Functional links - further exploratory work would be good here. Synthesis of the qualitative results will help provide ideas for other quantitative metrics. Social network datasets, especially Twitter or LinkedIn might be useful resources.

- Case studies - further case studies in each sector (e.g. digital health in London / Oxford / Nottingham, as well as Birmingham) would allow for richer sector insight and would give a clearer sense of specific cluster success factors / challenges. Case study sampling frames should be generated directly from the quantitative analysis, as has been done with the digital health example here.

References

Anderson, C. (2008). The End of Theory: How the data deluge makes the scientific method obsolete. Wired. **23 June**

Anyadike-Danes, M., K. Bonner, C. Drews and M. Hart (2013). Localisation of Industrial Activity across England's LEPs: 2008 and 20102. ERC Research Paper 15. Coventry, Aston University.

Arribas-Bel, D. (2014). "Accidental, open and everywhere: Emerging data sources for the understanding of cities." Applied Geography **49**: 45-53.

Bathelt, H., A. Malmberg and P. Maskell (2004). "Clusters and knowledge: local buzz, global pipelines and the process of knowledge creation." Progress in Human Geography **28**(1): 31-56.

Catini, R., D. Karamshuk, O. Penner and M. Riccaboni (2015). "Identifying geographic clusters: A network analytic approach." Research Policy **44**(9): 1749-1762.

Duranton, G. (2011). "California Dreamin': The feeble case for cluster policies." Review of Economic Analysis **3**(1): 3-45.

Duranton, G. and H. G. Overman (2005). "Testing for Localization Using Micro-Geographic Data." The Review of Economic Studies **72**(4): 1077-1106.

Einav, L. and J. D. Levin (2013). The Data Revolution and Economic Analysis. National Bureau of Economic Research Working Paper Series No. 19035. Cambridge, MA, NBER.

Garnsey, E. and P. Heffernan (2005). "High-technology clustering through spin-out and attraction: The Cambridge case." Regional Studies **39**(8): 1127-1144.

Hall, P. (2000). "Creative Cities and Economic Development." Urban Studies **37**(4): 639-649.

Hall, P. and K. Pain (2006). The Polycentric Metropolis: Learning from mega-city regions in Europe. London, Earthscan.

Jacobs, J. (1969). The Economy of Cities. London, Vintage.

Kerr, W. and S. Kominers (2015). "Agglomerative Forces and Cluster Shapes." Review of Economics and Statistics **97**(4): 877-899.

Marshall, A. (1918). Principles of Economics. New York, Macmillan.

Martin, R. and P. Sunley (2003). "Deconstructing clusters: chaotic concept or policy panacea?" Journal of Economic Geography **3**(1): 5-35.

Mateos-Garcia, J., H. Bakhshi and M. Lenel (2014). A Map of the UK Games Industry. London, NESTA.

Nathan, M. and H. Overman (2013). "Agglomeration, clusters, and industrial policy." Oxford Review of Economic Policy **29**(2): 383-404.

Nathan, M. and A. Rosso (2015). "Mapping digital businesses with Big Data: some early findings from the UK " Research Policy **44**(9): 1714-1733.

Nathan, M. and E. Vandore (2014). "Here be startups: exploring London's 'Tech City' digital cluster." Environment and Planning A **46**(10): 2283-2299.

Porter, M. (2003). "The Economic Performance of Regions." Regional Studies **37**(6-7): 545-546.

Saxenian, A.-L. (2006). The New Argonauts: Regional Advantage in a Global Economy. Cambridge, MA, Harvard University Press.

Scott, A. (1988). New industrial spaces: Flexible production organization and regional development in North America and Western Europe. London, Pion.

Storper, M. (1997). The Regional World: Territorial Development in a Global Economy. New York, Guilford.

Tech City UK (2015). Tech Nation. London, TCUK.

Varian, H. R. (2014). "Big Data: New Tricks for Econometrics." Journal of Economic Perspectives **28**(2): 3-28.

This publication available from www.gov.uk/beis

Contacts us if you have any enquiries about this publication, including requests for alternative formats, at:

Department for Business, Energy and Industrial Strategy
1 Victoria Street
London SW1H 0ET
Tel: 020 7215 5000

Email: enquiries@beis.gov.uk